

# Frames and numerical approximation

Ben Adcock\*

Daan Huybrechs†

December 15, 2016

## Abstract

Functions of one or more variables are usually approximated with a basis; a complete, linearly independent set of functions that spans an appropriate function space. The topic of this paper is the numerical approximation of functions using the more general notion of frames; that is, complete systems that are generally redundant but provide stable infinite representations. While frames are well-known tools in image and signal processing, coding theory and other areas of applied mathematics, their use in numerical analysis is far less widespread. Yet, as we show via example, frames are more flexible than bases, and can be constructed easily in a range of problems where finding orthonormal bases with desirable properties (rapid convergence, high resolution power, etc) is difficult or impossible. Examples and references given in this paper indicate that frames already appear in a variety of existing numerical methods, although they are not often identified as such.

A major difficulty in computing best approximations in such systems is that frames necessarily lead to ill-conditioned linear systems of equations. The ill-conditioned regime is often avoided in applications, or perceived as a disadvantage. However, we show that frame approximations can in fact be computed numerically up to an error of order  $\sqrt{\epsilon}$  with a simple algorithm, where  $\epsilon$  is a threshold parameter that can be chosen close to machine precision. Moreover, this accuracy can be improved to order  $\epsilon$  with modifications to the algorithm. Crucially, the order of convergence down to this limit is determined by the existence of representations of the function being approximated that are accurate and have small-norm coefficients. We demonstrate the existence of such representations in all our examples. Overall, our analysis suggests that frames are a natural generalization of bases in which to develop numerical approximation. In particular, even in the presence of severe ill-conditioning, the frame condition imposes sufficient mathematical structure on the redundant set in order to give rise to good approximations in finite precision calculations.

**Keywords** frames, function approximation, ill-conditioning, singular value decomposition

**AMS subject classifications** 42C15, 42C30, 41A10, 65T40

## 1 Introduction

Frames are a generalization of orthogonal and Riesz bases. They are indispensable tools in modern signal and image processing, and are used widely in a range of other problems, such as compression, source coding, robust transmission and sampling theory [12, 32, 51, 52, 62]. Yet frames are not so

---

\*Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A 1S6, Canada (ben\_adcock@sfu.ca, <http://www.benadcock.ca>)

†Department of Computer Science, University of Leuven, Celestijnenlaan 200A, BE-3001 Leuven, Belgium (daan.huybrechs@cs.kuleuven.be, <http://people.cs.kuleuven.be/~daan.huybrechs/>)

well known in numerical analysis. Although they arise in approximation problems in a number of ways, a systematic and general study of numerical frame approximation does not appear to have been undertaken.

The purpose of this paper is consider frame approximations from this perspective. By means of motivation, we introduce three classes of problems in numerical computing where frames already occur naturally, or where they may potentially lead to better methods. Our main objective is to examine the extent to which frame approximations are accurate and numerically stable, and the properties of a frame which ensure both.

A central theme of this paper is the difference between the behaviour of infinite frames and the corresponding truncated frames used in approximation. Unlike for orthonormal bases, stability of truncated frames does not follow from the properties of infinite frames. Indeed, the linear systems associated with frame approximations are necessarily ill-conditioned. The surprising fact, however, is that the numerically-stable frame approximations are still possible in spite of this ill-conditioning, provided these systems are properly regularized. Crucially, this means that, unlike in the case of orthonormal bases, there are key difference between ‘theoretical’ frame approximations (e.g. the best approximation) and ‘numerical’ frame approximations (i.e. the solution of the regularized system). Understanding and documenting these differences is a key aspect of this paper.

## 1.1 Orthonormal bases

Suppose that  $\Phi = \{\phi_n\}_{n \in \mathbb{N}}$  is an orthonormal basis of a separable Hilbert space  $H$ . Two key properties of  $\Phi$  are the straightforward representation of  $f$  in the basis using the inner product on  $H$ ,

$$f = \sum_{n \in \mathbb{N}} \langle f, \phi_n \rangle \phi_n, \quad \forall f \in H, \quad (1.1)$$

where the infinite sum converges in  $H$ , and Parseval’s identity

$$\|f\|^2 = \sum_{n \in I} |\langle f, \phi_n \rangle|^2, \quad \forall f \in H. \quad (1.2)$$

If the *coefficients*  $\langle f, \phi_n \rangle$  are known (or have been computed), approximation in  $\Phi$  is a straightforward affair. One simply replaces (1.1) by a finite expansion

$$f \approx \sum_{n=1}^N \langle f, \phi_n \rangle \phi_n. \quad (1.3)$$

This approximation has the beneficial property of being the orthogonal projection onto the space  $H_N = \text{span}\{\phi_n\}_{n=1}^N$ , and therefore the best approximation to  $f$  from  $H_N$  in the norm of  $H$ .

## 1.2 Frames

A set  $\Phi = \{\phi_n\}_{n \in \mathbb{N}}$  is called a *frame* for  $H$  if the span of  $\Phi$  is dense in  $H$  and  $\Phi$  satisfies the so-called *frame condition*

$$A\|f\|^2 \leq \sum_{n \in \mathbb{N}} |\langle f, \phi_n \rangle|^2 \leq B\|f\|^2, \quad \forall f \in H, \quad (1.4)$$

for constants  $A, B > 0$ . The optimal constants  $A, B > 0$  such that (1.4) holds, i.e. the largest possible  $A$  and the smallest possible  $B$ , are referred to as the *frame bounds* [32, 38]. We recall the theory of frames in more detail in §2.

Generalizing Parseval's identity, the frame condition expresses a norm equivalence between the  $\ell^2$ -norm of the coefficients  $\{\langle f, \phi_n \rangle\}_{n \in \mathbb{N}}$  and the Hilbert space norm of  $f$ . Yet frames differ from orthonormal bases in a number of key respects:

- (i) The frame elements  $\phi_n$  are not generally orthogonal.
- (ii) A frame is typically *redundant*. That is, any  $f \in H$  may have more than one expansion  $f = \sum_{n \in \mathbb{N}} c_n \phi_n$  with  $c = \{c_n\}_{n \in \mathbb{N}} \in \ell^2(\mathbb{N})$ .
- (iii) Unless the frame is *tight* (see §2), a representation such as (1.1) does not hold.

### 1.3 Computing orthogonal projections with frames

While (i) means that frames are more flexible than orthonormal bases (indeed, (1.4) is far less restrictive a condition than orthogonality), it presents an immediate difficulty for numerical approximation with frames. Even if the coefficients  $\langle f, \phi_n \rangle$  are available, the orthogonal projection onto  $H_N = \text{span}\{\phi_1, \dots, \phi_N\}$  does not have an explicit expression such as (1.3) and thus requires a computation. In the notation of our paper, this equates to solving the linear system

$$G_N x = y, \quad y = \{\langle f, \phi_n \rangle\}_{n=1}^N, \quad (1.5)$$

where  $G_N$  is the  $N \times N$  *truncated Gram matrix*

$$G_N = \{\langle \phi_m, \phi_n \rangle\}_{n,m=1}^N \in \mathbb{C}^{N \times N}.$$

If  $x \in \mathbb{C}^N$  is the solution of (1.5), the orthogonal projection is given by  $\sum_{n=1}^N x_n \phi_n$ .

There is also a second practical issue in frame computations, which stems from (ii). Due to their orthogonality, approximations with orthonormal bases are inherently stable ( $G_N = I$  for orthonormal bases). It is tempting to think that the frame condition (1.4) endows frame approximations with a similar stability. However, while a subset  $\Phi_N = \{\phi_n\}_{n=1}^N$  of a frame is indeed a frame for its span  $H_N$ , and thus satisfies a frame condition, its frame bounds  $A_N$  and  $B_N$  may behave wildly as  $N \rightarrow \infty$ , even when the infinite frame bounds  $A$  and  $B$  are mild. We shall see examples later in this paper where the ratio  $B_N/A_N$  grows superalgebraically fast in  $N$ . This instability of truncated frames is equivalent to ill-conditioning of the Gram matrix  $G_N$ , which stems from the noninvertibility of the Gram operator  $\mathcal{G}$  of the infinite frame (such noninvertibility is due to (ii)). Approximating  $\mathcal{G}$  by the finite matrix  $G_N$  results in small, nonzero eigenvalues and hence ill-conditioning [46]. Understanding this ill-conditioning and its effect on the resulting numerical frame approximation obtained by solving a regularized version of (1.5) is the central theme of this paper.

### 1.4 Motivations

Orthonormal bases are ubiquitous in numerical analysis. Important cases include Fourier and Chebyshev bases, in which case fast algorithms exist to (approximately) compute the expansions based on interpolation [72]. A major disadvantage of orthogonal bases however is their inflexibility. To illustrate, consider the problem of approximating smooth functions of one or more variables. While it is easy to construct good<sup>1</sup> orthogonal bases of functions on simple domains such as intervals, it is much harder to do so in higher dimensions unless the domain is particularly simple (e.g. a

---

<sup>1</sup>The word ‘good’ in this paper is taken to mean *spectrally* convergent, i.e. having rates of convergence depending only on the smoothness of the function being approximated.

hypercube). It is also problematic to find a good basis for singular functions, or to force periodicity on nonperiodic problems in order to take advantage of the FFT.

In this paper we show that good frames can be easily found for all these problems. In particular, we identify a simple frame with spectral rates of convergence for approximating functions defined on arbitrary Lipschitz domains. These examples illustrate three different generic constructions which always lead to frames: namely, restrictions of orthonormal bases to subdomains, augmentation of an orthonormal basis by a finite number of additional terms, and concatenation of several orthonormal bases. This leads us to opine that frames may be useful tools for many problems in numerical analysis where constructing orthonormal bases is difficult or impossible.

## 1.5 Overview and main results

We restrict our focus in this paper to two key properties: the convergence of finite approximations in a frame and their stability. We shall mostly ignore the question of efficiency, since this is highly dependent on the type of frame used and in this paper we strive for generality. We will return to this topic briefly in §7.

Our main conclusion is the following. In spite of the extreme ill-conditioning of the linear system (1.5), accurate frame approximations, in a sense we make precise below, can be computed numerically. The linear system (1.5) is regularized using a simple truncated Singular Value Decomposition (SVD) of the Gram matrix. This is not necessarily the best or most efficient algorithm, but its analysis is well suited to illustrate the issues involved in numerical frame approximations.

We recall the main elements in the theory of frames in §2. In §3 we introduce three generic constructions of frames that are useful in numerical approximations along with our three main examples. All these examples deal with the problem of approximating functions where it is not straightforward or even desirable to use orthonormal bases, and where frame approximations present a viable alternative.

Our analysis commences in §4. The stability of an infinite frame expansion is determined by the ratio  $B/A$  of the frame bounds. However, unlike for orthogonal or Riesz bases, passing from the countable frame  $\Phi$  to a finite subset  $\Phi_N = \{\phi_n\}_{n=1}^N$  necessarily causes a deterioration in the frame bounds. We document this phenomenon in §4. Lemma 4.2 shows how frame bounds deteriorate after truncation, and Proposition 4.3 establishes that the effect can be arbitrarily bad. The condition numbers  $\kappa(G_N)$  for the three example frames are estimated in §4.3. Each exhibits algebraic, superalgebraic or exponential growth in  $N$ .

We consider the computation of the best approximation via orthogonal projection in §5. We first show in Proposition 5.1 that the  $\ell^2$ -norm of the *exact* solution vector  $x$  to the system (1.5) is generally unbounded in  $N$ , due to the ill-conditioning. Hence, computing  $x$  with any accuracy for large  $N$  is impossible. However, the situation improves markedly after regularizing  $G_N$  by truncating its SVD below a threshold  $\epsilon$ . In Theorem 5.3 we show that the convergence of the regularized projection to a function  $f$  is dictated by how well  $f$  can be approximated by vectors of coefficients with small norm. Specifically, if  $\mathcal{P}_N^\epsilon f$  is the regularized projection then

$$\|f - \mathcal{P}_N^\epsilon f\| \leq \|f - \mathcal{T}_N z\| + \sqrt{\epsilon}\|z\|, \quad \forall z \in \mathbb{C}^N, f \in \mathbf{H}, \quad (1.6)$$

where  $\mathcal{T}_N z = \sum_{n=1}^N z_n \phi_n \in \mathbf{H}_N$ . The first term in the right-hand side is standard and denotes the approximation error corresponding to a coefficient vector  $z$ . The second term is uncommon in the literature on frames, and indeed it is specific to a numerical frame approximation: it regularizes the approximation by penalizing the  $\ell^2$ -norm of  $z$  weighted by  $\sqrt{\epsilon}$ . The estimate implies that the approximation error will be small if  $f$  can be well represented in the frame (first term) with coefficients of small norm (second term). The existence of such representations in the first place is

guaranteed by the frame condition, and that is why we argue that the mathematical structure of a frame seems a highly appropriate general context to discuss function approximation in redundant systems.

Theorem 5.4 accordingly shows that the solution vector of the regularized system eventually (for large  $N$ ) exhibits small norm, though there may be an initial regime in which it is large. The precise result is

$$\|x^\epsilon\| \leq 1/\sqrt{\epsilon}\|f - \mathcal{T}_N z\| + \|z\|, \quad \forall z \in \mathbb{C}^N. \quad (1.7)$$

The solution may initially be large due to the  $1/\sqrt{\epsilon}$  in the first term in the right hand side, but this term goes to zero as  $N$  increases. Thus, the computation of the regularized solution has bounded instability, despite the unbounded ill-conditioning of the Gram matrix. The price to pay for this beneficial property is that the true convergence rate of the best approximation may not be realized after regularization. Instead, one finds best approximations subject to having a small-norm coefficient vector. In practice, these solutions are often more desirable since they are inherently stable. The details depend on the frame at hand, and are described in §5.4 for our example frames.

Finally, a point of clarification. The reader may be tempted to conclude from this discussion that frame approximations are of limited use in practice, since they can obtain at best  $\mathcal{O}(\sqrt{\epsilon})$  accuracy. We caution that is not the case. In §6 we will briefly describe a generalized frame approximation which achieves  $\mathcal{O}(\epsilon)$  accuracy, and thus genuine numerical stability. The full analysis of these techniques (which builds on this paper) will be described in an upcoming work [8].

## 1.6 Relation to existing work

Frames were introduced in the context of nonharmonic Fourier series by Duffin & Schaeffer [38]. They were later developed by Daubechies, Grossmann and Meyer [36] in the 1980's with the systematic study of wavelets. Since then they have become an integral part of modern signal and image processing, compression, coding theory and sampling theory. For overviews, see [12, 22, 26, 32, 35, 51, 52].

In frame theory, approaches to numerical frame approximations – with the notable exception of [46, 70] – have usually centred around either explicitly identifying an appropriate dual frame or by numerically inverting the frame operator (equivalent to computing the canonical dual frame) [23, 24, 25, 29, 30, 31, 33, 34, 68]. We refer to §2.3 for further details. While these approaches are useful for approximations with, for example, Gabor or wavelet frames and their various generalizations (e.g. multi-wavelets [32], ridgelets [19], curvelets [20, 62] and shearlets [54]), for the problems which motivate this paper the dual frame expansion usually converges too slowly to be of practical use. We give several examples of this phenomenon later. On the other hand, our focus in this paper is on computing best approximations with frames, or more precisely, surrogates obtained from solving regularized systems. The regularized Gram systems we consider in this paper have previously been studied in [46, 70] in the context of frames of exponentials arising in nonuniform sampling problems. In particular, [46, Thm. 5.17] asserts convergence of the coefficients  $x^\epsilon$  to the so-called frame coefficients (see §2.3) as  $N \rightarrow \infty$ . We develop this result by establishing the convergence rate (1.6) and finite stability bound (1.7) for arbitrary frames.

Our study of numerical frame approximations stems from previous works of the authors on so-called *Fourier extensions* [16, 18], also known as *Fourier continuation* or *Fourier embedding* in the context of numerical PDEs [65]. The connection to frame theory was first explored in [49], and further developed in the one-dimensional setting in [9]. A by-product of this paper is an extension of [9] to  $d \geq 1$  dimensions. Yet we stress that our main results apply to any frame, not just Fourier extensions.

We also draw several interesting connections to other fields. In particular, the disparity between truncated frames and infinite frames is related to the spectral theory of self-adjoint operators, and specifically the phenomenon of pollution in the finite section method [37, 56]. We also make links to the topic of time- and band-limiting [48], in particular the prolate spheroidal wavefunctions [67], and to classical regularization theory of ill-posed problems [40, 45, 64].

## 2 Preliminaries

### 2.1 Orthogonal and Riesz bases

For the remainder of the paper,  $\Phi = \{\phi_n\}_{n \in I}$  denotes a subset of a separable Hilbert space  $H$  over the field  $\mathbb{C}$ , where  $I$  is a countable index set. We write  $\langle \cdot, \cdot \rangle$  and  $\|\cdot\|$  for the inner product and norm on  $H$  respectively. The set  $\Phi$  is an *orthonormal* basis for  $H$  if  $\text{span}(\Phi)$  is dense in  $H$  and  $\langle \phi_n, \phi_m \rangle = \delta_{n,m}, \forall n, m \in I$ . Recall that orthonormal bases satisfy Parseval's identity

$$\|x\| = \left\| \sum_{n \in I} x_n \phi_n \right\|, \quad \forall x = \{x_n\}_{n \in I} \in \ell^2(I). \quad (2.1)$$

Here, and throughout this paper, we use  $\|x\|$  to denote the  $\ell^2$ -norm of a (finite or infinite) sequence  $x$ . Equivalently,

$$\|f\|^2 = \sum_{n \in I} |\langle f, \phi_n \rangle|^2, \quad \forall f \in H. \quad (2.2)$$

A *Riesz* basis is a generalization of an orthonormal basis. Such a basis is no longer orthogonal, but it satisfies the following relaxed version of Parseval's identity:

$$A\|x\|^2 \leq \left\| \sum_{n \in I} x_n \phi_n \right\|^2 \leq B\|x\|^2, \quad \forall x = \{x_n\}_{n \in I} \in \ell^2(I), \quad (2.3)$$

where  $A, B > 0$  are positive constants. Throughout this paper, whenever constants  $A$  and  $B$  are introduced in an inequality such as this, they will be taken to be the optimal constants such that the corresponding inequality holds. Note that this inequality also implies the following relaxed version of (2.2):

$$A\|f\|^2 \leq \sum_{n \in I} |\langle f, \phi_n \rangle|^2 \leq B\|f\|^2, \quad \forall f \in H. \quad (2.4)$$

A popular example of a Riesz basis are the hat functions often used in finite element methods. They are not orthogonal, but they are a basis for their span. Another example are more general B-splines [32].

Any Riesz basis has a unique *dual* Riesz basis  $\Psi$ . This basis satisfies

$$\langle \phi_n, \psi_m \rangle = \delta_{n,m}, \quad n, m \in I.$$

For this reason, the Riesz basis and its dual are sometimes called *biorthogonal*. Any function  $f \in H$  has a unique representation in the Riesz basis given explicitly in terms of inner products with the dual basis:

$$f = \sum_{n \in I} \langle f, \psi_n \rangle \phi_n. \quad (2.5)$$

Here, equality is taken to mean convergence in the norm of  $H$ . Note that an orthonormal basis is self-dual, i.e.  $\Psi = \Phi$ .

## 2.2 Frames

A countable system  $\Phi$  is called a *frame* for  $H$  if  $\text{span}(\Phi)$  is dense in  $H$  and if  $\Phi$  satisfies the *frame condition*, which we repeat here for convenience:

$$A\|f\|^2 \leq \sum_{n \in I} |\langle f, \phi_n \rangle|^2 \leq B\|f\|^2, \quad \forall f \in H. \quad (2.6)$$

It follows from the Parseval equality (2.2) and its generalization (2.4) that orthonormal and Riesz bases are frames. However, most frames are not bases. Indeed, frames are generally not  $\omega$ -independent: that is, there exist nonzero sequences of coefficients  $x \in \ell^2(I)$  with  $\sum_{n \in I} x_n \phi_n = 0$  [32, Sec. 5.5]. Conversely, bases are always  $\omega$ -independent. As mentioned, this *redundancy* gives frames far greater flexibility than bases, making them substantially easier to construct for particular problems.

We now introduce some further standard notions pertaining to frames. Associated to any frame  $\Phi$  is the so-called *synthesis* operator

$$\mathcal{T} : \ell^2(I) \rightarrow H, \quad y = \{y_n\}_{n \in I} \mapsto \sum_{n \in I} y_n \phi_n,$$

which maps a sequence to an expansion in the frame. Its adjoint, the *analysis* operator, is given by

$$\mathcal{T}^* : H \rightarrow \ell^2(I), \quad f \mapsto \{\langle f, \phi_n \rangle\}_{n \in I},$$

and the composition  $\mathcal{S} = \mathcal{T}\mathcal{T}^*$  is known as the *frame operator*:

$$\mathcal{S} : H \rightarrow H, \quad f \mapsto \sum_{n \in I} \langle f, \phi_n \rangle \phi_n.$$

For an orthonormal basis  $\mathcal{S}f$  converges to  $f$ , but for a Riesz basis and a more general frame this is no longer the case in general. Still, the frame operator is a useful object. It is self-adjoint by construction, and it follows from the frame condition that  $\mathcal{S}$  is also bounded and invertible on  $H$  [32, Lemma 5.1.5], and satisfies  $A\mathcal{I} \leq \mathcal{S} \leq B\mathcal{I}$ , where  $\mathcal{I}$  is the identity operator on  $H$  and  $A$  and  $B$  are the frame bounds.

The Gram operator of a frame is defined by  $\mathcal{G} = \mathcal{T}^*\mathcal{T}$ . That is,

$$\mathcal{G} : \ell^2(\mathbb{N}) \rightarrow \ell^2(\mathbb{N}), \quad \mathcal{G}x = \left\{ \sum_{m \in I} \langle \phi_m, \phi_n \rangle x_m \right\}_{n \in I}. \quad (2.7)$$

Note that  $\mathcal{G}$  is a bounded operator on  $\ell^2(\mathbb{N})$ , but is not in general invertible (see §4). We may also view  $\mathcal{G}$  as the infinite matrix  $G = \{\langle \phi_n, \phi_m \rangle\}_{n,m \in I}$ . Throughout this paper all infinite matrices are equivalent to bounded operators on  $\ell^2(\mathbb{N})$ .

A frame is said to be *tight* if  $A = B$ , in which case  $\mathcal{S} = A\mathcal{I}$  is a multiple of the identity. However,  $\mathcal{G}$  does not have this property unless the frame is also an orthonormal basis.

We shall also need two further notions. First, a frame is said to be *exact* if it ceases to be a frame when any one element is removed. A frame that is not exact is referred to as *inexact*. Second, we say a frame  $\{\phi_n\}_{n \in I}$  is *linearly independent* if every finite subset  $\{\phi_n\}_{n \in J}$ ,  $|J| < \infty$ , is linearly independent.

A frame is exact if and only if it is a Riesz basis [32, Theorem 5.5.4]. Hence, for the remainder of this paper we will assume that all frames are inexact. We shall also assume that all frames are linearly independent. This is mainly for convenience, and it will be the case in all examples discussed. Note that a linearly independent frame is not necessarily a Riesz basis. See [32, Chpt. 6] for further discussion on independence and the relations between frames and Riesz bases.

## 2.3 Dual frames

A frame  $\Psi = \{\psi_n\}_{n \in I} \subseteq H$  is called a dual frame for  $\Phi$  if

$$f = \sum_{n \in I} \langle f, \psi_n \rangle \phi_n = \sum_{n \in I} \langle f, \phi_n \rangle \psi_n, \quad \forall f \in H. \quad (2.8)$$

An inexact frame necessarily has more than one dual frame. Moreover, a frame and its duals are not biorthogonal, unlike the case of Riesz bases. However, there is a unique so-called *canonical dual frame*  $\Psi = \{\psi_n\}_{n \in I}$ , given by  $\psi_n = \mathcal{S}^{-1} \phi_n$ , where  $\mathcal{S}$  is the frame operator. Since  $\Psi$  is a dual frame, one has

$$f = \sum_{n \in I} \langle f, \mathcal{S}^{-1} \phi_n \rangle \phi_n = \sum_{n \in I} \langle \mathcal{S}^{-1} f, \phi_n \rangle \phi_n. \quad (2.9)$$

Moreover, the dual frame bounds are  $1/B$  and  $1/A$  respectively, and it follows that

$$1/B \|f\|^2 \leq \sum_{n \in I} |\langle f, \mathcal{S}^{-1} \phi_n \rangle|^2 \leq 1/A \|f\|^2, \quad \forall f \in H. \quad (2.10)$$

We refer to the coefficients  $a = \{\langle f, \mathcal{S}^{-1} \phi_n \rangle\}_{n \in I}$  as the *frame coefficients* of  $f$ . Note that these coefficients have the beneficial property that, amongst all possible representations of  $f$  in  $\Phi$ , they have the smallest norm. If  $f = \sum_{n \in I} a_n \phi_n = \sum_{n \in I} c_n \phi_n$ , where  $a$  are the frame coefficients, then  $\|c\| \geq \|a\|$  [32, Lem. 5.4.2].

At this stage, one might be tempted to approximate  $f$  by computing its dual frame coefficients and truncating the expansion (2.8). This could potentially be done either by analytically identifying a dual frame (when possible) or by numerically inverting the frame operator [23, 24, 25, 29, 30, 31, 33, 34]. However, computational issues aside (in the case where the frame is not tight, computing the canonical dual frame is nontrivial as it requires inversion of an operator, i.e.  $\mathcal{S}$ , with infinite-dimensional domain and range) the approximation  $\sum_{n \in I_N} \langle \mathcal{S}^{-1} f, \phi_n \rangle \phi_n$  is generally not the orthogonal projection onto  $H_N = \text{span}\{\phi_n : n \in I_N\}$ . For the examples which motivate this paper, this expansion typically converges much more slowly than the orthogonal projection.<sup>2</sup> See §3 and Figs. 3 and 4 for several examples of this phenomenon.

## 2.4 Truncated frames

For each  $N \in \mathbb{N}$  we introduce the truncated systems  $\Phi_N = \{\phi_n\}_{n \in I_N}$  where  $I_N \subseteq I$  and  $|I_N| = N$ . For simplicity, we assume that the index sets  $\{I_N\}_{N \in \mathbb{N}}$  are nested and satisfy

$$I_1 \subseteq I_2 \subseteq \dots \subseteq I, \quad \bigcup_{N=1}^{\infty} I_N = I. \quad (2.11)$$

The system  $\Phi_N$  is a frame for its span  $H_N = \text{span}(\Phi_N)$ . We write  $A_N, B_N > 0$  for the frame bounds, so that

$$A_N \|f\|^2 \leq \sum_{n \in I_N} |\langle f, \phi_n \rangle|^2 \leq B_N \|f\|^2, \quad \forall f \in H_N, \quad (2.12)$$

---

<sup>2</sup>This is in contrast to the case of wavelet frames and their various generalizations, which are specifically designed to have accurate dual frame representations.



and let

$$\begin{aligned}
\mathcal{T}_N : \mathbb{C}^N &\rightarrow H_N, \quad y = \{y_n\}_{n \in I_N} \mapsto \sum_{n \in I_N} y_n \phi_n, \\
\mathcal{T}_N^* : H_N &\rightarrow \mathbb{C}^N, \quad f \mapsto \{\langle f, \phi_n \rangle\}_{n \in I_N}, \\
\mathcal{S}_N = \mathcal{T}_N \mathcal{T}_N^* : H_N &\rightarrow H_N, \quad f \mapsto \sum_{n \in I_N} \langle f, \phi_n \rangle \phi_n,
\end{aligned} \tag{2.13}$$

be the truncated analysis, synthesis and frame operators respectively. We also define the truncated Gram operator  $\mathcal{G}_N = \mathcal{T}_N^* \mathcal{T}_N$  and the associated  $N \times N$  Gram matrix

$$G_N = \{\langle \phi_m, \phi_n \rangle\}_{n,m \in I_N} \in \mathbb{C}^{N \times N}. \tag{2.14}$$

Since the frame  $\Phi$  is linearly independent by assumption, it follows that the Gram matrix  $G_N$  is full rank. Indeed, if  $x = \{x_n\}_{n \in I_N} \in \mathbb{C}^N$  then  $x^* G_N x = \left\| \sum_{n \in I_N} x_n \phi_n \right\|^2$  and, by linear independence, the right-hand side is zero if and only if  $x = 0$ .

## 2.5 Best approximations and rates of convergence

Given a frame  $\Phi$  and a finite subset  $\Phi_N$  a key task is to compute the orthogonal projection  $\mathcal{P}_N$  onto  $H_N = \text{span}(\Phi_N)$ . Observe that  $\mathcal{P}_N f$  is the best approximation to  $f$  from  $H_N$  in  $\|\cdot\|$ . For  $f \in H$ , write

$$\mathcal{P}_N f = \sum_{n \in I_N} x_n \phi_n, \quad x = \{x_n\}_{n \in I_N} \in \mathbb{C}^N. \tag{2.15}$$

Since it is defined by the orthogonality conditions  $\langle \mathcal{P}_N f, \phi_n \rangle = \langle f, \phi_n \rangle$ ,  $\forall n \in I_N$ , one has that the coefficients  $x = \{x_n\}_{n \in I_N}$  are the unique solution of the linear system

$$G_N x = y, \quad y = \{\langle f, \phi_n \rangle\}_{n \in I_N}. \tag{2.16}$$

Hence, computing the best approximation in a frame requires solving an  $N \times N$  linear system. This turns out to be ill-conditioned, which is in direct contrast with the case of an orthonormal basis, wherein the Gram matrix  $G_N$  is the identity and  $x_n = y_n = \langle f, \phi_n \rangle$ . As we shall see, the major difficulty that arises when computing with frames is having to solve ill-conditioned systems such as (2.16).

Besides stability, a primary concern of this paper is the convergence rate of approximations such as  $\mathcal{P}_N f$ . To this end, we distinguish three types of convergence of an approximation  $f_N$  to a function  $f$ . First, we say that  $f_N$  converges *algebraically* fast to  $f$  at rate  $k$  if  $\|f - f_N\| = \mathcal{O}(N^{-k})$  as  $N \rightarrow \infty$ . Second, if  $\|f - f_N\|$  decays faster than any algebraic power of  $N^{-1}$  then we say that  $f_N$  converges *superalgebraically* fast to  $f$ . Third, we say that  $f_N$  converges *geometrically* fast to  $f$  if there exists a  $\rho > 1$  such that  $\|f - f_N\| = \mathcal{O}(\rho^{-N})$ .

## 3 Examples of frames

We now introduce three examples of frames that will be used throughout the paper to interpret the general results proved later for arbitrary frames. Each example illustrates the flexibility gained in numerical approximations by allowing redundancy in the approximation.

**Example 1. Fourier frames for complex geometries.** Let  $\Omega \subseteq \mathbb{R}^d$  be a compact domain and  $f : \Omega \rightarrow \mathbb{R}$  a smooth function. Besides simple domains (cubes, toruses, spheres, etc), it is in general

very difficult to find orthonormal bases for  $H = L^2(\Omega)$  with simple, explicit expressions and whose orthogonal projections exhibit spectral convergence. However, it is straightforward to find a frame with this property.

Since  $\Omega$  is compact, it can be contained in a hypercube. Without loss of generality, suppose that  $\Omega \subseteq (-1, 1)^d$ . Now consider a system of functions formed by the restriction of the orthonormal Fourier basis on  $(-1, 1)^d$  to  $\Omega$ :

$$\Phi = \{\phi_n\}_{n \in \mathbb{Z}^d}, \quad \phi_n(t) = 2^{-d/2} e^{i\pi n \cdot t}, \quad t \in \Omega. \quad (3.1)$$

This system is not an orthonormal basis of  $L^2(\Omega)$ , but it is a tight, linearly-independent frame with  $A = B = 1$ . If we introduce the truncated frames

$$\Phi_N = \{\phi_n\}_{n \in I_N}, \quad I_N = \left\{ n = (n_1, \dots, n_d) \in \mathbb{Z}^d : -\frac{1}{2}N^{1/d} \leq n_1, \dots, n_d < \frac{1}{2}N^{1/d} \right\}, \quad (3.2)$$

then the convergence rate of the corresponding orthogonal projections  $\mathcal{P}_N f$  is spectral: that is, algebraic for functions with finite smoothness and superalgebraic for smooth functions (Proposition 5.8).

The approximation based on the frame (3.1) is known as a Fourier *extension* (or *continuation*) [16, 18] in the one-dimensional case, and occasionally referred to as a Fourier *embedding* in higher dimensions [15, 65]. The connection to frames was first explored in [49], and further analysis of the one-dimensional case was given in [6, 9, 59].

Recalling the discussion in §2.3, this frame is an example where the canonical dual frame expansion (2.9) converges slowly. Indeed, since  $\Phi$  is tight it is its own canonical dual frame, and therefore the frame coefficients are  $a_n = \langle f, \phi_n \rangle$ . They are precisely the Fourier coefficients of the extension  $\tilde{f}$  of  $f$  by zero to  $(-1, 1)^d$ :

$$\langle f, \phi_n \rangle = \int_{\Omega} f(x) \phi_n(x) dx = \int_{(-1, 1)^d} \tilde{f}(x) \phi_n(x) dx.$$

As a result, the expansion (2.9) is nothing more than the Fourier series of  $\tilde{f}$  restricted to  $\Omega$ . Unless  $f$  vanishes smoothly on the boundary  $\partial\Omega$ , this expansion converges slowly and suffers from a Gibbs-type phenomenon near  $\partial\Omega$ . In contrast, the convergence of  $\mathcal{P}_N f$  is spectral, regardless of the shape of  $\Omega$ .

**Remark 3.1** This example illustrates a general principle: the restriction of a Riesz basis on a domain  $\Omega_e$  to a subset  $\Omega \subset \Omega_e$  always results in a frame. If the basis on  $\Omega_e$  is orthonormal, the corresponding frame on  $\Omega$  is tight. Such a construction has been used for the numerical solution of PDEs in complex geometries. Recent examples of embedding methods implicitly based on such ‘extension’ frames include [58, 69]. See also [11, 17, 61] for a method based on one-dimensional extensions.

**Example 2. Augmented Fourier basis.** Consider the case of smooth, nonperiodic functions  $f : [-1, 1] \rightarrow \mathbb{R}$ . Polynomial bases have good convergence properties for such functions but relatively bad resolution power for oscillatory functions. On the other hand, the Fourier approximation of a nonperiodic function suffers from the Gibbs phenomenon at  $t = \pm 1$  and converges only slowly in the  $L^2$ -norm. One way to seek to remedy this situation is to augment the Fourier basis by a finite number  $K \in \mathbb{N}$  of additional functions  $\psi_1, \dots, \psi_K$ , leading to the system

$$\Phi = \{\varphi_n\}_{n \in \mathbb{Z}} \cup \{\psi_k\}_{k=1}^K, \quad \varphi_n(t) = \frac{1}{\sqrt{2}} e^{i\pi n t}. \quad (3.3)$$

To save unnecessary generalizations, we will assume that  $\psi_k = \sqrt{k+1/2}P_k$ , where  $P_k \in \mathbb{P}_k$  is the  $k^{\text{th}}$  Legendre polynomial. Note that  $\{\psi_k\}_{k=1}^K$  is an orthonormal basis for the space

$$\mathbb{P}_K^0 = \left\{ p \in \mathbb{P}_K : \int_{-1}^1 p(t) dt = 0 \right\}.$$

Since  $\{\phi_n\}_{n \in \mathbb{Z}}$  is an orthonormal basis and  $K$  is finite,  $\Phi$  forms a frame for  $H = L^2(-1, 1)$  with frame bounds  $A = 1$  and  $B = 2$ . It is also linearly independent, since no finite sum of the complex exponentials  $\phi_n$  is exactly equal to a nonconstant algebraic polynomial. If

$$\Phi_N = \{\varphi_n : n = -\frac{N-K}{2}, \dots, \frac{N-K}{2} - 1\} \cup \{\psi_k\}_{k=1}^K, \quad N \geq K, \quad N - K \text{ even},$$

is the truncated frame (we will not consider the odd case, although it presents few difficulties), then orthogonal projections with respect to this frame inherit the optimal resolution properties of the Fourier basis, yet converge algebraically with rate  $K$  for all sufficiently smooth functions (Proposition 5.9). Conversely, the canonical dual frame expansion converges at roughly the same rate as the Fourier expansion of  $f$  (see Proposition SM2.2 of the supplementary material [7]).

The idea of augmenting the Fourier basis with a finite number of additional functions is an old one, arguably dating back to Krylov [53]. These functions endow the frame with good approximation properties by implicitly subtracting the jump discontinuities of  $f$  at the interval endpoints. This smoothed function now has a faster converging Fourier expansion, leading to the better convergence stated above. This approach is also commonly referred to as Eckhoff's method [39] or Euler–MacLaurin interpolants [50]. Whilst the convergence of this approximation has been extensively studied (see [1, 2] and references therein), the connection with frame theory is, to the best of our knowledge, new.

**Remark 3.2** An overall principle illustrated by this example is that adding a finite set of elements of  $H$  to a Riesz (in particular, orthogonal) basis always results in a frame; a so-called *Riesz frame* [32, Sec. 6.2]. Although we focus on polynomials enhancing the Fourier basis here, augmenting a basis with additional terms to incorporate features of the function to be approximated (in this case, smoothness) is quite a general idea. Other examples might include piecewise polynomial functions in the presence of interior discontinuities, or compactly-supported functions in the case of local variations such as oscillations.

**Example 3. Polynomial plus modified polynomials.** Several problems in numerical analysis call for the approximations of functions of the form

$$f(t) = w(t)g(t) + h(t), \quad t \in [-1, 1], \quad (3.4)$$

where  $g, h$  are smooth functions of  $t$  but  $w \in L^\infty(-1, 1)$  may be singular, oscillatory or possessing some other kind of feature which makes approximation difficult. The presence of  $w(t)$  usually means that the polynomial approximation of  $f$  converges only slowly in  $N$ . One particular instance is

$$w(t) = (1+t)^\alpha, \quad 0 < \alpha < 1, \quad (3.5)$$

which corresponds to a weak endpoint singularity of the function  $f$ . Note that (3.5) has been considered in [27, 28] in the context of PDEs with endpoint singularities. For other examples corresponding to an oscillatory function  $w(t)$ , see [47, 66].

If  $w(t)$  is known explicitly or has been estimated accurately (as it is in the applications mentioned above), then it is natural to use it to construct a frame to approximate  $f$ . Let  $\{\varphi_n\}_{n \in \mathbb{N}}$  be the

orthonormal basis of Legendre polynomials (we could also use Chebyshev polynomials here, but we shall use Legendre for simplicity). Then we form the system

$$\Phi = \{\varphi_n\}_{n \in \mathbb{N}} \cup \{\psi_n\}_{n \in \mathbb{N}}, \quad \psi_n(t) = w(t)\varphi_n(t). \quad (3.6)$$

Since  $w \in L^\infty(-1, 1)$  and  $\{\varphi_n\}_{n \in \mathbb{N}}$  is an orthonormal basis, this system gives rise to a frame for the space  $H = L^2(-1, 1)$ . A simple calculation gives that

$$A = 1 + \operatorname{ess\,inf}_{t \in (-1, 1)} |w(t)|^2, \quad B = 1 + \operatorname{ess\,sup}_{t \in (-1, 1)} |w(t)|^2.$$

This frame is linearly dependent if and only if  $w$  is a rational function of two polynomials. We assume from now on that this is not the case. For even  $N$ , we define the truncated frames by  $\Phi_N = \{\varphi_n\}_{n=1}^{N/2} \cup \{\psi_n\}_{n=1}^{N/2}$ . Orthogonal projections with respect to this frame are spectrally convergent with respect to the smoothness of  $g$  and  $h$  (Proposition 5.10). Conversely, the convergence of the dual frame expansion is generally not spectral, but algebraic at a fixed rate (Proposition SM3.2).

Note that more terms can be included in (3.4), i.e.  $f(t) = \sum_{i=1}^K w_i(t)g_i(t)$  for functions  $w_1, \dots, w_K$ . If these are known, then this would lead to the frame construction  $\Phi = \cup_{i=1}^K \{\psi_{i,n}\}_{n \in \mathbb{N}}$ , where  $\psi_{i,n}(t) = w_i(t)\varphi_n(t)$ . For simplicity, we consider only the case (3.4), although the generalization is conceptually straightforward. As in Example 2, the interpretation of this approach as a frame approximation has not, to the best of our knowledge, been considered before.

**Remark 3.3** The composition of a finite number of Riesz or orthonormal bases always results in a frame. More generally, the composition of several frames is still a frame. We note that the concept of concatenations of bases or frames is widely-used in signal and image processing [21]. Typically, images and signals may have substantially sparser representations in the resulting frame than in a single orthonormal basis, which yields benefits in tasks such as compression and denoising [62].<sup>3</sup>

## 4 Truncated Gram matrices and ill-conditioning

Since we have assumed linear independence, a truncated frame is a Riesz basis for its span. Hence the truncated frame bounds  $A_N$  and  $B_N$  are the same as the Riesz bounds. However, this finite basis is very skewed, and this results in the familiar ill-conditioning of truncated frames. In this section we explore the effect of this truncation in more detail.

### 4.1 The Gram operator

We recall some properties of the Gram operator  $\mathcal{G}$  of a frame  $\Phi$ . The Gram operator is a self-adjoint, nonnegative operator on  $l^2(I)$  with closed range. It is bounded, and its restriction  $\mathcal{G} : l^2(I) \rightarrow \operatorname{Ran}(\mathcal{G})$  is invertible. Its spectrum  $\sigma(\mathcal{G})$  satisfies

$$\{B\} \subseteq \sigma(\mathcal{G}) \subseteq \{0\} \cup [A, B],$$

where  $A, B$  are the frame bounds [46]. As shown in [71],  $\mathcal{G}$  is compact if and only if  $H$  is finite-dimensional, and  $\mathcal{G}$  is positive if and only if  $\Phi$  is a Riesz basis. Hence, in this paper  $\mathcal{G}$  is singular

<sup>3</sup>For completeness, we should note that the system  $\{\psi_n\}_{n \in \mathbb{N}}$  is only a Riesz basis if  $w(t)$  is bounded away from zero on  $[-1, 1]$ . This is not the case in (3.5), for example. However,  $\{\psi_n\}_{n \in \mathbb{N}}$  is always a *Bessel sequence* (see, for example, [32, Def. 3.2.2]); that is, a sequence for which the upper frame condition  $\sum_{n \in \mathbb{N}} |\langle f, \psi_n \rangle|^2 \leq B \|f\|^2$  holds, but for which the lower frame condition need not hold. The composition of a Riesz basis or frame with a finite number of Bessel sequences is also always a frame.

and thus  $\text{Ker}(\mathcal{G}) \neq \{0\}$ . Nonetheless, since  $\mathcal{G}$  has closed range, we may define its Moore–Penrose pseudoinverse  $\mathcal{G}^\dagger : \ell^2(I) \rightarrow \ell^2(I)$  [42, 71]. One then has the following relation between  $\mathcal{G}$ ,  $\mathcal{G}^\dagger$  and the frame bounds:

$$A = \|\mathcal{G}^\dagger\|^{-1}, \quad B = \|\mathcal{G}\|. \quad (4.1)$$

Here  $\|\cdot\|$  is the operator norm on  $\ell^2(I)$ .

## 4.2 Truncated Gram matrices

We now consider conditioning of the matrix  $G_N$ . From the above discussion, we immediately note the following:

**Lemma 4.1.** *The truncated Gram matrix  $G_N$  of a linearly-independent frame  $\Phi$  is invertible with*

$$\|G_N^{-1}\|^{-1} = A_N, \quad \|G_N\| = B_N,$$

where  $A_N$  and  $B_N$  are the frame bounds of the truncated frame  $\Phi_N$ . In particular, its condition number

$$\kappa(G_N) = \|G_N\| \|G_N^{-1}\| = B_N/A_N,$$

is equal to the ratio of the truncated frame bounds.

In practice, we will also use the following characterization of the frame bounds:

$$A_N = \min_{\substack{x \in \mathbb{C}^N \\ \|x\|=1}} \|\mathcal{T}_N x\|^2, \quad B_N = \max_{\substack{x \in \mathbb{C}^N \\ \|x\|=1}} \|\mathcal{T}_N x\|^2, \quad (4.2)$$

which follows immediately from the fact that  $G_N = \mathcal{T}_N^* \mathcal{T}_N$ . This characterization lends itself to an intuitive interpretation. The constant  $A_N$  measures how small in norm an element of  $H_N$  can be, while having unit discrete norm of its expansion coefficients in the truncated frame. Equivalently, it measures how well the zero element  $0 \in H_N$  can be approximated by an element  $H_N$  with unit-norm coefficients. It is clear from this that when frame elements are close to being linearly dependent, the constant  $A_N$  can be quite small. On the other hand,  $B_N$  measures how large a function in  $H$  can be with bounded coefficients. Here, one expects that  $B_N$  remains bounded even for nearly dependent frame elements.

**Lemma 4.2.** *Let  $\Phi$  be a linearly-independent frame. Then*

- (i) *the sequences  $\{A_N\}_{N \in \mathbb{N}}$  and  $\{B_N\}_{N \in \mathbb{N}}$  are monotonically nonincreasing and nondecreasing respectively,*
- (ii)  *$B_N \leq B$  for all  $N$  and  $B_N \rightarrow B$  as  $N \rightarrow \infty$ ,*
- (iii)  *$\inf_N A_N > 0$  if and only if  $\Phi$  is a Riesz basis.*

*Proof.* Part (i) follows immediately from (2.11) and (4.2), as does the observation that  $B_N \leq B$  in part (ii). To deduce convergence, let  $0 < \epsilon < \sqrt{B}$  be arbitrary and suppose that  $x \in \ell^2(I)$ ,  $\|x\| = 1$ , is such that  $\sqrt{B} \geq \|\mathcal{T}x\| = \sqrt{\langle \mathcal{S}x, x \rangle} \geq \sqrt{B} - \epsilon$ . Let  $z \in \mathbb{C}^N$  be such that  $z_n = x_n$  for  $n \in I_N$  and suppose that  $x^N \in \ell^2(I)$  is the extension of  $z$  by zero. Then

$$\sqrt{B_N} \geq \frac{\|\mathcal{T}_N z\|}{\|z\|} = \frac{\|\mathcal{T}x^N\|}{\|x^N\|} \geq \frac{\|\mathcal{T}x\| - \|\mathcal{T}(x - x^N)\|}{\|x\| + \|x - x^N\|} \geq \frac{\sqrt{B} - \epsilon - \sqrt{B}\|x - x^N\|}{1 + \|x - x^N\|}.$$

Since  $x^N \rightarrow x$  as  $N \rightarrow \infty$ , we deduce that  $\sqrt{B} \geq \sqrt{B_N} \geq \sqrt{B} - 2\epsilon$  for all sufficiently large  $N$ . Since  $\epsilon$  was arbitrary we see that  $B_N \rightarrow B$ . Finally, for part (iii) we use [32, Prop. 6.1.2].  $\square$

This lemma implies that the truncated Gram matrices  $G_N$  are necessarily ill-conditioned for large  $N$ . Such ill-conditioning can also be arbitrarily bad:

**Proposition 4.3.** *Let  $\{\phi_n\}_{n \in \mathbb{N}}$  be an orthonormal basis of  $\mathcal{H}$  and let  $g \in \mathcal{H}$ ,  $\|g\| = 1$ , be such that  $\langle g, \phi_n \rangle \neq 0$  for infinitely many  $n$ . Then the system  $\Phi^g = \{g, \phi_1, \phi_2, \dots\}$  is a linearly-independent frame for  $\mathcal{H}$  with bounds  $A = 1$  and  $B = 2$ . Moreover, if  $\Phi_N^g = \{g, \phi_1, \phi_2, \dots, \phi_{N-1}\}$ , then the finite frame bounds are given by*

$$A_N = 1 - \sqrt{\sum_{n=1}^{N-1} |\langle g, \phi_n \rangle|^2}, \quad B_N = 1 + \sqrt{\sum_{n=1}^{N-1} |\langle g, \phi_n \rangle|^2}. \quad (4.3)$$

*Proof.* Certainly  $\text{span}(\Phi^g)$  is dense in  $\mathcal{H}$  since it contains an orthonormal basis. Now let  $f \in \mathcal{H}$  be arbitrary and note that  $|\langle f, g \rangle|^2 + \sum_{n \in \mathbb{N}} |\langle f, \phi_n \rangle|^2 = |\langle f, g \rangle|^2 + \|f\|^2$ . It follows that  $\Phi^g$  is a frame with  $A = 1$  and  $B = 2$ . Moreover, since  $\langle g, \phi_n \rangle \neq 0$  for infinitely-many  $n$ ,  $g$  cannot be written as a finite sum of the  $\phi_n$ 's. Hence  $\Phi^g$  is linearly independent.

Consider the truncated frame  $\Phi_N^g$ . First note that (4.3) holds trivially if  $\langle g, \phi_n \rangle = 0$  for  $n = 1, \dots, N-1$ . Therefore we may assume that  $\langle g, \phi_n \rangle \neq 0$  for some  $n = 1, \dots, N-1$ . Let  $x = \{x_n\}_{n=0}^{N-1}$  be an eigenvector of  $G_N$  with eigenvalue  $\lambda$ . Then

$$x_0 + \sum_{n=1}^{N-1} x_n \overline{\langle g, \phi_n \rangle} = \lambda x_0, \quad \langle g, \phi_m \rangle x_0 + x_m = \lambda x_m, \quad m = 1, \dots, N-1.$$

Suppose first that  $\lambda = 1$ . Then  $x_0 = 0$  and  $x_1, \dots, x_{N-1}$  satisfy  $\sum_{n=1}^{N-1} x_n \overline{\langle g, \phi_n \rangle} = 0$ . Hence  $\lambda = 1$  is an eigenvalue of multiplicity  $N-2$ . Now consider the case  $\lambda \neq 1$ . Then  $x_m = (\lambda - 1)^{-1} \langle g, \phi_m \rangle x_0$ ,  $m = 1, \dots, N-1$ , and therefore

$$x_0 + x_0(\lambda - 1)^{-1} \sum_{n=1}^{N-1} |\langle g, \phi_n \rangle|^2 = \lambda x_0.$$

Note that  $x_0 \neq 0$  since  $x \neq 0$ , and therefore  $\lambda^2 - 2\lambda + 1 - \sum_{n=1}^{N-1} |\langle g, \phi_n \rangle|^2 = 0$ . The roots of this equation are precisely (4.3). Counting multiplicities, we see that these roots must be the maximal and minimal eigenvalues of  $G_N$  respectively.  $\square$

Since  $\|g\| = 1$  for this frame, we have  $A_N \leq \sqrt{\sum_{n \geq N} |\langle g, \phi_n \rangle|^2}$ . Hence the decay of  $A_N$  is related to the decay of the coefficients  $\langle g, \phi_n \rangle$  in the basis  $\{\phi_n\}_{n \in \mathbb{N}}$ . The better  $g$  is represented in this basis, the worse the conditioning of the truncated Gram matrix. This result illustrates how easy it is for the truncated Gram matrices to be ill-conditioned: we can create arbitrarily bad conditioning merely by adding one additional element to an orthonormal basis. When more elements are added (as in Example 2) or a whole orthonormal basis is added (as in Example 3), it is not surprising that the corresponding Gram matrices can be exceedingly ill-conditioned. See §4.3 for further details.

**Remark 4.4** The truncated Gram matrices  $G_N$  are equivalent to the  $N \times N$  finite sections of the infinite Gram matrix  $G$ . Recall that  $\sigma(G) \subseteq \{0\} \cup [A, B]$  and that  $0 \in \sigma(G)$ . Unfortunately, as illustrated in Figure 1 for the three examples,  $\sigma(G_N)$  does not lie within  $\{0\} \cup [A, B]$ . Instead, spurious eigenvalues are introduced in the spectral gap between 0 and the lower frame bound  $A$ ; a well-known phenomenon referred to as spectral pollution in the finite section method [37, 56]. Of particular relevance to this paper are the small eigenvalues of  $G_N$ . These translate into ill-conditioning of  $G_N$ , and hence the need for regularization. Note that  $0 \notin \sigma(G_N)$  for any  $N$  since the frame is linearly independent, yet as seen in Lemma 4.1, small eigenvalues necessarily occur.

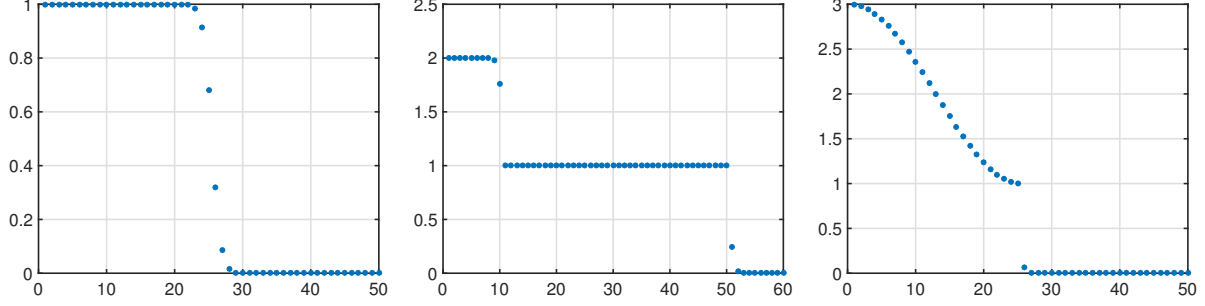


Figure 1: The eigenvalues of  $G_{50}$  for Examples 1–3 (left to right). The parameters used were  $T = 2$  (Example 1),  $K = 10$  (Example 2) and  $w(t) = \sqrt{1+t}$  (Example 3).

### 4.3 Examples

We now discuss  $\kappa(G_N)$  for the examples introduced in §3. Proofs of the results in this section are given in the supplementary material [7].

**Example 1.** If  $\Omega = (-\frac{1}{T}, \frac{1}{T})$  is an interval, where  $T > 1$ , then  $\kappa(G_N) = \mathcal{O}(E(T)^N)$  as  $N \rightarrow \infty$ , where  $E(T) = \cot^2(\pi/(4T)) > 1$  [9]. Hence the condition number is geometrically large in  $N$  – see Figure 2(a). A similar, albeit somewhat weaker, result also holds in arbitrary dimensions:

**Proposition 4.5.** *Let  $\Omega \subseteq (-1, 1)^d$  be a Lipschitz domain and consider the frame (3.1). Then the condition numbers  $\kappa(G_N)$  grow superalgebraically fast in  $N$ .*

The explanation of this result is rather simple. One can show that the kernel  $\text{Ker}(\mathcal{G})$ , a subset of  $\ell^2(I)$ , consists precisely of the sequences of Fourier coefficients of functions on  $(-1, 1)^d$  which vanish on  $\Omega$  (Proposition SM1.1 of the supplementary material [7]). Now consider a smooth function  $g$  with this property. Then its Fourier expansion on  $(-1, 1)^d$  converges superalgebraically fast to  $g$  on  $(-1, 1)^d$ , and therefore to zero on the domain  $\Omega$ . The ratio of the norm of the truncated Fourier series on  $\Omega$  divided by its norm on  $(-1, 1)^d$  is an upper bound for  $A_N$ , implying ill-conditioning at a superalgebraic rate.

**Example 2.** In this case,  $\kappa(G_N)$  grows algebraically fast at a rate depending on  $K$ :

**Proposition 4.6.** *Let  $K \in \mathbb{N}$  be fixed and consider the frame (3.3). Then  $\kappa(G_N) \gtrsim N^{2K-1}$  as  $N \rightarrow \infty$ .*

The intuition behind this result is as follows. One can show that the kernel of the gram operator  $\mathcal{G}$  has dimension  $K$ , and consists of infinite vectors  $x \in \ell^2(\mathbb{Z})$  which are comprised of the coefficients  $\{\langle p, \psi_k \rangle\}_{k=1}^K$  and  $\langle -p, \varphi_n \rangle\}_{n \in \mathbb{Z}}$ , where  $p$  is an arbitrary polynomial in  $\mathbb{P}_K^0$ . See Proposition SM2.1. It is possible to construct a polynomial  $p \in \mathbb{P}_K^0$  which has  $K$  orders of periodic smoothness (of course,  $p$  is analytic, but it is not periodic in general). This function has Fourier coefficients  $\langle p, \varphi_n \rangle$  which decay like  $|n|^{-K-1}$  as  $n \rightarrow \pm\infty$ . Hence there is a function in  $H_N$ , i.e. the difference between  $p$  and its partial Fourier series, which is of magnitude  $\mathcal{O}(N^{-K})$  but which has  $\mathcal{O}(1)$  coefficients in the frame  $\Phi_N$ .

**Example 3.** In this case, we have the following :

**Proposition 4.7.** *Let  $\Phi$  be the frame (3.6) with  $w(t)$  given by (3.5). Then  $\kappa(G_N) \gtrsim 4^N$  as  $N \rightarrow \infty$  up to an algebraic factor in  $N$ .*

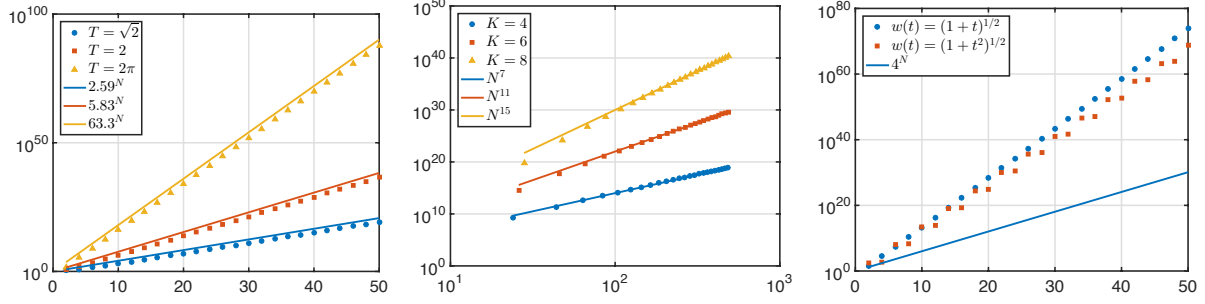


Figure 2: The condition numbers  $\kappa(G_N)$  for Examples 1–3 (left to right) with parameters  $T = \sqrt{2}, 2, 2\pi$ ,  $K = 4, 6, 8$  and  $w(t) = \sqrt{1+t}$  respectively. The solid lines show the bounds in Propositions 4.5–4.7. Computations were carried out in *Mathematica* using additional precision.

The idea of this result is similar to that of Example 2 (see Proposition SM3.1 for a description of  $\text{Ker}(\mathcal{G})$  in this case). We choose a polynomial  $q(t) = (1+t)^{N/2-1}$  such that  $w(t)q(t)$  has several orders of smoothness at  $t = 0$  in spite of the algebraic singularity there. Hence, it can be well approximated by a single polynomial  $p(t)$ . The difference  $p - wq \in \mathcal{H}_N$  has  $\mathcal{O}(1)$  coefficients, yet its norm is very small which implies ill-conditioning of the Gram matrix.

Numerical illustrations of these three estimates are shown in Figure 2. Unlike in Examples 1 & 2, the lower bound of  $4^N$  in Example 3 does not give a good estimate of the true growth of  $\kappa(G_N)$ .

## 5 Computing best approximations

We now consider the computation of the orthogonal projection  $\mathcal{P}_N f$ . Our first assertion is that it is impossible in general to compute  $\mathcal{P}_N f$ , since the coefficients of this approximation can grow rapidly with  $N$ . However, with a simple numerical scheme one can compute best approximations subject to having small norm coefficients. This solves two problems: the numerical solution is computable, and since it has bounded coefficients it is also more stable.

### 5.1 Impossibility of computing best approximations

Computing  $\mathcal{P}_N f = \sum_{n \in I_N} x_n \phi_n$  requires solving the ill-conditioned linear system (2.16). If  $x = \{x_n\}_{n \in I_N}$  and  $y = \{\langle f, \phi_n \rangle\}_{n \in I_N}$  then by Lemma 4.1 we have

$$\|x\| = \|G_N^{-1}y\| \leq \|G_N^{-1}\| \|y\| \leq A_N^{-1} \sqrt{B} \|f\|.$$

Hence, the coefficients  $x$  of the orthogonal projection may, in the worst case, grow as rapidly as  $A_N^{-1}$ . Of course, this is only an upper bound, and therefore may not be achieved for a fixed  $f \in \mathcal{H}$ . However, it is easy to create an example where the growth of  $\|x\|$  is the same order as that of  $A_N^{-1}$ .

**Proposition 5.1.** *Let  $\{\phi_n\}_{n \in \mathbb{N}}$ ,  $g \in \mathcal{H}$ ,  $\Phi^g$  and  $\Phi_N^g$  be as in Proposition 4.3 and suppose that  $f \in \mathcal{H}$ ,  $\|f\| = 1$  is given by*

$$f = \frac{\sqrt{6}}{\pi} \sum_{n \in \mathbb{N}} \frac{\text{sign}(\langle g, \phi_n \rangle)}{n} \phi_n,$$



where, for  $\omega \in \mathbb{C}$ ,  $\text{sign}(\omega) = \omega/|\omega|$  if  $\omega \neq 0$  and  $\text{sign}(\omega) = 0$  otherwise. Suppose also that  $\sup_{n \in \mathbb{N}} |\langle g, \phi_n \rangle| n^2 < \infty$ . If  $x = \{x_n\}_{n \in I_N}$  is the solution of (2.16) then

$$\|x\| \geq \sqrt{\frac{\sum_{n \geq N} n^{-2} |\langle g, \phi_n \rangle|}{\sum_{n \geq N} |\langle g, \phi_n \rangle|^2}} \geq \left( \frac{\sqrt{\pi}}{6} \max_{n \geq N} \{n |\langle g, \phi_n \rangle|\} \right)^{-1}.$$

*Proof.* The orthogonality of the basis  $\{\phi_n\}_{n \in \mathbb{N}}$  and the fact that  $\|g\| = 1$  means that the system  $G_N x = y$  is equivalent to

$$x_0 + \sum_{m=1}^{N-1} \overline{\langle g, \phi_m \rangle} x_m = \langle f, g \rangle, \quad \langle g, \phi_n \rangle x_0 + x_n = \langle f, \phi_n \rangle, \quad n = 1, \dots, N-1.$$

Substituting the second equation into the first gives

$$x_0 \left( 1 - \sum_{m=1}^{N-1} |\langle g, \phi_m \rangle|^2 \right) + \sum_{m=1}^{N-1} \langle f, \phi_m \rangle \overline{\langle g, \phi_m \rangle} = \langle f, g \rangle,$$

and after rearranging and using the definition of  $f$ , this gives

$$x_0 = \frac{\sum_{m \geq N} \langle f, \phi_m \rangle \overline{\langle g, \phi_m \rangle}}{\sum_{m \geq N} |\langle g, \phi_m \rangle|^2} = \frac{\sqrt{6} \sum_{m \geq N} n^{-1} |\langle g, \phi_m \rangle|}{\pi \sum_{m \geq N} |\langle g, \phi_m \rangle|^2}. \quad (5.1)$$

Observe that

$$\sum_{m \geq N} |\langle g, \phi_m \rangle|^2 \leq \max_{n \geq N} \{n |\langle g, \phi_n \rangle|\} \sum_{m \geq N} m^{-1} |\langle g, \phi_m \rangle|.$$

Substituting this into (5.1) and noting that  $\|x\| \geq |x_0|$  now gives the result.  $\square$

Given a basis  $\{\phi_n\}$  suppose we augment it with an element  $g \in \mathcal{H}$  that is well approximated in the basis; for example,  $|\langle g, \phi_n \rangle| = \mathcal{O}(n^{-\alpha-1})$  for some  $\alpha \geq 0$ . Then  $A_N^{-1} \gtrsim N^{\alpha+1/2}$  by Proposition 4.3 and, if  $x$  is as in Proposition 5.1, then  $\|x\| \gtrsim N^\alpha$ . Hence, there exists coefficients  $x$  of a fixed function  $f$  which grow almost as fast as the condition number of the Gram matrix  $G_N$ .

Although this example is synthetic, it illustrates the general principle that the coefficients of the orthogonal projection  $\mathcal{P}_N f$  in a truncated frame approximation can grow at a similar rate to that of the condition number. Hence it is generally impossible to compute these coefficients accurately in finite-precision arithmetic. To see this in a more practical setting, in Table 1 we display the coefficients for Example 1 when applied to several different functions. As is evident, only for the entire function  $f(x) = \exp(x)$  is the growth of the coefficients avoided. For the other two functions, which are less smooth, we witness geometric growth of the coefficients, mirroring that of the condition number (see Proposition 4.5).

$N$	10	20	40	80	160
$f(t) = \exp(t)$	1.77e0	1.81e0	1.84e0	1.86e0	1.87e0
$f(t) = \frac{1}{1+16t^2}$	2.27e0	5.05e1	3.64e4	2.32e10	1.13e22
$f(t) =  t ^5$	2.12e-1	3.67e-1	1.76e4	7.62e26	6.09e91
$\kappa(G_N)$	1.84e6	5.64e13	8.01e28	2.35e59	2.90e120

Table 1: The  $\ell^2$ -norm  $\|x\|$  of the coefficients of the orthogonal projection  $\mathcal{P}_N f$  for Example 1.

## 5.2 Truncated SVD projections

To regularize the ill-conditioned system (2.16), we resort to a familiar approach: compute the SVD of  $G_N$ , discard all singular values below a tolerance  $\epsilon$  and then find the solution  $x^\epsilon$  of the resulting system. The entries of the vector  $x^\epsilon$  are no longer the coefficients of the orthogonal projection but rather the projection onto a smaller space  $H_N^\epsilon$  depending on  $\epsilon$ . Despite having discarded many of the singular values, as we now discuss this projection can still approximate  $f$  to high accuracy.

We first require some notation. Since  $G_N$  is positive definite its singular values  $\sigma_1, \dots, \sigma_N$  are its eigenvalues and its SVD takes the form

$$G_N = V \Sigma V^*,$$

where  $V \in \mathbb{C}^{N \times N}$  is unitary and  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_N)$  is diagonal. Write  $\{v_n : n \in I_N\}$  for the columns of  $V$ , which are the left/right singular vectors of  $G_N$ , i.e.  $G_N v_n = \sigma_n v_n$ ,  $n \in I_N$ . To each singular vector we associate a element  $\xi_n$  of  $H_N$ :

$$\xi_n = \sum_{m \in I_N} (v_n)_m \phi_m = \mathcal{T}_N v_n \in H_N.$$

It follows from the orthogonality of  $v_n$  that the functions  $\xi_n$  are orthogonal in  $H$ :

$$\langle \xi_n, \xi_m \rangle = \langle \mathcal{T}_N v_n, \mathcal{T}_N v_m \rangle = \langle v_n, \mathcal{T}_N^* \mathcal{T}_N v_m \rangle = \sigma_m \langle v_n, v_m \rangle = \sigma_m \delta_{n,m}. \quad (5.2)$$

As a result they form an orthogonal basis for  $H_N$ .

Given a tolerance  $\epsilon > 0$ , let  $\Sigma^\epsilon$  be the diagonal matrix with  $n^{\text{th}}$  entry  $\sigma_n$  if  $\sigma_n > \epsilon$  and zero otherwise. Define

$$G_N^\epsilon = V \Sigma^\epsilon V^*. \quad (5.3)$$

Then the truncated SVD coefficients  $x^\epsilon$  are  $x^\epsilon = (G_N^\epsilon)^\dagger y = V(\Sigma^\epsilon)^\dagger V^* y$ , where  $\dagger$  denotes the pseudoinverse. Note that  $(\Sigma^\epsilon)^\dagger$  is diagonal with  $n^{\text{th}}$  entry equal to  $1/\sigma_n$  if  $\sigma_n > \epsilon$  and 0 otherwise. We may also write  $x$  and  $x^\epsilon$  as follows:

$$x = \sum_{n \in I_N} \frac{\langle y, v_n \rangle}{\sigma_n} v_n, \quad x^\epsilon = \sum_{\sigma_n > \epsilon} \frac{\langle y, v_n \rangle}{\sigma_n} v_n, \quad (5.4)$$

where  $\langle \cdot, \cdot \rangle$  denotes the Euclidean inner product on  $\mathbb{C}^N$ . Given  $x^\epsilon$ , much like with the coefficients  $x$ , we define the approximation

$$\mathcal{P}_N^\epsilon f = \mathcal{T}_N x^\epsilon = \sum_{n \in I_N} (x^\epsilon)_n \phi_n.$$

Observe that  $\langle y, v_n \rangle = \sum_{m \in I_N} \langle f, \phi_m \rangle \overline{(v_n)_m} = \langle f, \xi_n \rangle$ , and therefore

$$\mathcal{P}_N^\epsilon f = \sum_{\sigma_n > \epsilon} \frac{\langle f, \xi_n \rangle}{\sigma_n} \xi_n, \quad \mathcal{P}_N f = \sum_{n \in I_N} \frac{\langle f, \xi_n \rangle}{\sigma_n} \xi_n. \quad (5.5)$$

Thus,  $\mathcal{P}_N^\epsilon$  is the orthogonal projection from  $H$  to  $H_N^\epsilon$ , where  $H_N^\epsilon = \text{span}\{\xi_n : \sigma_n > \epsilon\}$ .

**Remark 5.2** Note that  $\|\xi_n\| = \sqrt{\sigma_n}$  due (5.2). Hence any singular vector  $v_n$  with small singular value  $\sigma_n$  also corresponds to a function  $\xi_n$  that has small norm in  $H$ . In other words, these functions do not contribute much to the approximation. From this point of view, it is not all that surprising that they can be discarded.

In Example 1, the singular vectors  $v_n$  and the functions  $\xi_n$  correspond precisely to the so-called *prolate spheroidal wave sequences* and *prolate spheroidal wave functions*, introduced by Slepian, Landau and Pollak in the study of bandlimited extrapolation [55, 67]. These are a central object of study in the subfield of harmonic analysis and signal processing that focuses on time-frequency localization of signals [35, 48]. In our setting, the prolate functions corresponding to small singular values are small on  $\Omega$  but large on the extrapolated region  $(-1, 1)^d \setminus \Omega$ , i.e. they are approximately supported away from  $\Omega$  and hence they do not influence the approximation substantially [63].

### 5.3 Analysis of the truncated SVD projection $\mathcal{P}_N^\epsilon$

We now consider the error of the projection  $\mathcal{P}_N^\epsilon$ . The following is our main result:

**Theorem 5.3.** *The truncated SVD projection  $\mathcal{P}_N^\epsilon$  satisfies*

$$\|f - \mathcal{P}_N^\epsilon f\| \leq \|f - \mathcal{T}_N z\| + \sqrt{\epsilon}\|z\|, \quad \forall z \in \mathbb{C}^N, f \in \mathcal{H}. \quad (5.6)$$

*Proof.* Let  $z \in \mathbb{C}^N$ . Since  $\mathcal{P}_N^\epsilon$  is the orthogonal projection onto  $\mathcal{H}_N^\epsilon$ , we have

$$\|f - \mathcal{P}_N^\epsilon f\| \leq \|f - \mathcal{P}_N^\epsilon \mathcal{T}_N z\| \leq \|f - \mathcal{T}_N z\| + \|\mathcal{T}_N z - \mathcal{P}_N^\epsilon \mathcal{T}_N z\|.$$

Note that  $\mathcal{T}_N z = \mathcal{P}_N \mathcal{T}_N z$  since  $\mathcal{T}_N z \in \mathcal{H}_N$ . Hence (5.5) and the orthogonality of the  $\xi_n$ 's gives

$$\|\mathcal{T}_N z - \mathcal{P}_N^\epsilon \mathcal{T}_N z\|^2 = \left\| \sum_{\sigma_n < \epsilon} \frac{\langle \mathcal{T}_N z, \xi_n \rangle}{\sigma_n} \xi_n \right\|^2 = \sum_{\sigma_n < \epsilon} \frac{|\langle \mathcal{T}_N z, \xi_n \rangle|^2}{\sigma_n}.$$

Observe that  $\langle \mathcal{T}_N z, \xi_n \rangle = \langle \mathcal{T}_N z, \mathcal{T}_N v_n \rangle = \langle z, G_N v_n \rangle = \sigma_n \langle z, v_n \rangle$  and therefore

$$\|\mathcal{T}_N z - \mathcal{P}_N^\epsilon \mathcal{T}_N z\|^2 = \sum_{\sigma_n < \epsilon} \sigma_n |\langle z, v_n \rangle|^2 < \epsilon \sum_{n \in I_N} |\langle z, v_n \rangle|^2 = \epsilon \|z\|^2,$$

where in the last step we use the fact that the vectors  $\{v_n\}_{n \in I_N}$  are orthonormal.  $\square$

This theorem establishes the claim made earlier in the paper: the convergence of the projection  $\mathcal{P}_N^\epsilon f$  is dictated by how well  $f$  can be approximated by coefficients  $z$  with small norm. Note that this situation is markedly different to the case of the projection  $\mathcal{P}_N f$ , wherein the analogous error bound is simply  $\|f - \mathcal{P}_N f\| \leq \|f - \mathcal{T}_N z\|$ ,  $\forall z \in \mathbb{C}^N$ . As we discuss in §5.4, the appearance of the term  $\sqrt{\epsilon}\|z\|$  can change the behaviour of the error in a key way.

Having analyzed the projection  $\mathcal{P}_N^\epsilon f$ , we now consider the behaviour of the coefficients  $x^\epsilon$ :

**Theorem 5.4.** *Let  $a = \{\langle f, \mathcal{S}^{-1} \phi_n \rangle\}_{n \in I}$  be the frame coefficients of  $f \in \mathcal{H}$ . Then the coefficients  $x^\epsilon$  of the truncated SVD projection  $\mathcal{P}_N^\epsilon$  satisfy*

$$\|x^\epsilon\| \leq 1/\sqrt{\epsilon} \|f - \mathcal{T}_N z\| + \|z\|, \quad \forall z \in \mathbb{C}^N, \quad (5.7)$$

and, if  $a^{N, \epsilon} \in \ell^2(I)$  is the extension of  $x^\epsilon$  by zero,

$$\|a - a^{N, \epsilon}\| \leq \left(1 + \sqrt{B/\epsilon}\right) \sqrt{\sum_{n \in I \setminus I_N} |a_n|^2} + \sqrt{\epsilon/A} \|a\|. \quad (5.8)$$

*Proof.* For the first part, we use (5.4) to write

$$x^\epsilon = \sum_{\sigma_n > \epsilon} \frac{\langle f, \xi_n \rangle}{\sigma_n} v_n = \sum_{\sigma_n > \epsilon} \frac{\langle f - \mathcal{T}_N z, \xi_n \rangle}{\sigma_n} v_n + \sum_{\sigma_n > \epsilon} \frac{\langle \mathcal{T}_N z, \xi_n \rangle}{\sigma_n} v_n.$$

Consider the first term on the right-hand side. By (5.2) and (5.5) we have

$$\left\| \sum_{\sigma_n > \epsilon} \frac{\langle f - \mathcal{T}_N z, \xi_n \rangle}{\sigma_n} v_n \right\|^2 = \sum_{\sigma_n > \epsilon} \frac{|\langle f - \mathcal{T}_N z, \xi_n \rangle|^2}{\sigma_n^2} \leq \frac{1}{\epsilon} \|\mathcal{P}_N^\epsilon(f - \mathcal{T}_N z)\|^2 \leq \frac{1}{\epsilon} \|f - \mathcal{T}_N z\|^2.$$

For the second term, we first notice that  $\langle \mathcal{T}_N z, \xi_n \rangle = \sigma_n \langle z, v_n \rangle$ , and therefore

$$\left\| \sum_{\sigma_n > \epsilon} \frac{\langle \mathcal{T}_N z, \xi_n \rangle}{\sigma_n} v_n \right\|^2 = \sum_{\sigma_n > \epsilon} |\langle z, v_n \rangle|^2 \leq \|z\|^2.$$

Combining these two bounds now gives the first result. For the second result, we first let  $a^N \in \mathbb{C}^N$  be such that  $a_n^N = a_n$ ,  $n \in I_N$ . Then

$$\|a - a^{N, \epsilon}\| \leq \sqrt{\sum_{n \in I \setminus I_N} |a_n|^2} + \|a^N - x^\epsilon\|.$$

Hence it suffices to estimate  $\|a^N - x^\epsilon\|$ . For this, we first note that  $f = \mathcal{S}\mathcal{S}^{-1}f = \mathcal{S}_N\mathcal{S}^{-1}f + (\mathcal{S} - \mathcal{S}_N)\mathcal{S}^{-1}f$ . Since  $\mathcal{S}_N$  is self-adjoint and  $\mathcal{S}_N\xi_n = \mathcal{T}_N\mathcal{T}_N^*\mathcal{T}_N v_n = \sigma_n\mathcal{T}_N v_n = \sigma_n\xi_n$  we have

$$\langle f, \xi_n \rangle = \langle \mathcal{S}_N\mathcal{S}^{-1}f, \xi_n \rangle + \langle (\mathcal{S} - \mathcal{S}_N)\mathcal{S}^{-1}f, \xi_n \rangle = \sigma_n \langle \mathcal{S}^{-1}f, \xi_n \rangle + \langle (\mathcal{S} - \mathcal{S}_N)\mathcal{S}^{-1}f, \xi_n \rangle.$$

Therefore

$$x^\epsilon = \sum_{\sigma_n > \epsilon} \frac{\langle f, \xi_n \rangle}{\sigma_n} v_n = \sum_{\sigma_n > \epsilon} \langle \mathcal{S}^{-1}f, \xi_n \rangle v_n + \sum_{\sigma_n > \epsilon} \frac{1}{\sigma_n} \langle (\mathcal{S} - \mathcal{S}_N)\mathcal{S}^{-1}f, \xi_n \rangle v_n. \quad (5.9)$$

Conversely, since  $a^N = \mathcal{T}_N^*\mathcal{S}^{-1}f$  we have that  $\langle a^N, v_n \rangle = \langle \mathcal{S}^{-1}f, \mathcal{T}_N v_n \rangle = \langle \mathcal{S}^{-1}f, \xi_n \rangle$ . Hence

$$a^N = \sum_{n \in I_N} \langle a^N, v_n \rangle v_n = \sum_{n \in I_N} \langle \mathcal{S}^{-1}f, \xi_n \rangle v_n. \quad (5.10)$$

Combining (5.9) and (5.10) now gives

$$\|a^N - x^\epsilon\| \leq \left\| \sum_{\sigma_n \leq \epsilon} \langle \mathcal{S}^{-1}f, \xi_n \rangle v_n \right\| + \left\| \sum_{\sigma_n > \epsilon} \frac{1}{\sigma_n} \langle (\mathcal{S} - \mathcal{S}_N)\mathcal{S}^{-1}f, \xi_n \rangle v_n \right\|. \quad (5.11)$$

Consider the first term. By orthogonality

$$\left\| \sum_{\sigma_n \leq \epsilon} \langle \mathcal{S}^{-1}f, \xi_n \rangle v_n \right\|^2 \leq \epsilon \sum_{\sigma_n \leq \epsilon} \frac{1}{\sigma_n} |\langle \mathcal{S}^{-1}f, \xi_n \rangle|^2 \leq \epsilon \|\mathcal{S}^{-1}f\|^2 \leq \epsilon/A \|a\|^2, \quad (5.12)$$

Here the final inequality follows from (2.6). Now consider the second term of (5.11). We have

$$\begin{aligned} \left\| \sum_{\sigma_n > \epsilon} \frac{1}{\sigma_n} \langle (\mathcal{S} - \mathcal{S}_N) \mathcal{S}^{-1} f, \xi_n \rangle v_n \right\|^2 &= \sum_{\sigma_n > \epsilon} \frac{1}{\sigma_n^2} |\langle (\mathcal{S} - \mathcal{S}_N) \mathcal{S}^{-1} f, \xi_n \rangle|^2 \\ &\leq \frac{1}{\epsilon} \sum_{\sigma_n > \epsilon} \frac{1}{\sigma_n} |\langle (\mathcal{S} - \mathcal{S}_N) \mathcal{S}^{-1} f, \xi_n \rangle|^2 \\ &\leq \frac{1}{\epsilon} \|(\mathcal{S} - \mathcal{S}_N) \mathcal{S}^{-1} f\|^2. \end{aligned}$$

Now

$$\|(\mathcal{S} - \mathcal{S}_N) \mathcal{S}^{-1} f\| = \left\| \sum_{n \in I \setminus I_N} a_n \phi_n \right\| = \sup_{\substack{g \in \mathcal{H} \\ g \neq 0}} \left\{ \frac{|\sum_{n > N} a_n \overline{g, \phi_n}|}{\|g\|} \right\} \leq \sqrt{B} \sqrt{\sum_{n \in I \setminus I_N} |a_n|^2}, \quad (5.13)$$

and therefore

$$\left\| \sum_{\sigma_n > \epsilon} \frac{1}{\sigma_n} \langle (\mathcal{S} - \mathcal{S}_N) \mathcal{S}^{-1} f, \xi_n \rangle v_n \right\|^2 \leq B/\epsilon \sum_{n \in I \setminus I_N} |a_n|^2.$$

Substituting this and (5.12) into (5.11) gives the result.  $\square$

Theorems 5.3 and 5.4 show a rather surprising conclusion. Despite severe ill-conditioning of the Gram matrix, which led us to discard all of its singular values of size less than  $\epsilon$ , one still gets convergence of  $\mathcal{P}_N^\epsilon f$  to within  $\sqrt{\epsilon}$  of  $f$ . Moreover, although the coefficients  $x^\epsilon$  may initially grow large (due to the  $1/\sqrt{\epsilon}$  factor in (5.7)), they too eventually converge to within  $\sqrt{\epsilon}$  of the frame coefficients of  $f$ .

The underlying reason for this, as detailed in the following theorem, is that the condition number of the mapping  $y = \{\langle f, \phi_n \rangle\}_{n \in I_N} \mapsto \mathcal{P}_N^\epsilon f$  from the data  $y$  to the projection  $\mathcal{P}_N^\epsilon f$  is much smaller – precisely, on the order of  $1/\sqrt{\epsilon}$  as opposed to  $1/\epsilon$  – than the condition number of the mapping from  $y$  to the coefficients  $x^\epsilon$  (which is just the condition number of the SVD truncated Gram matrix  $G_N^\epsilon$ ; see (5.3)). Hence, whilst errors in the coefficients may be on the order of  $1/\epsilon$ , they result in much smaller errors in the projection  $\mathcal{P}_N^\epsilon$ .

**Theorem 5.5.** *The absolute condition number of the mapping  $\mathbb{C}^N \rightarrow \mathcal{H}_N^\epsilon$ ,  $y \mapsto \mathcal{T}_N(G_N^\epsilon)^\dagger y$  satisfies*

$$\kappa \leq \min \left\{ 1/\sqrt{\sigma_{\min}(G_N)}, 1/\sqrt{\epsilon} \right\}, \quad \forall N \in \mathbb{N},$$

where  $\sigma_{\min}(G_N)$  is the minimal singular value of the Gram matrix  $G_N$ .

*Proof.* By linearity, the condition number of the mapping  $y \mapsto \mathcal{T}_N(G_N^\epsilon)^\dagger y$  is  $\kappa = \max_{\substack{y \in \mathbb{C}^N \\ \|y\|=1}} \|\mathcal{T}_N(G_N^\epsilon)^\dagger y\|$ .

We have

$$\|\mathcal{T}_N(G_N^\epsilon)^\dagger y\|^2 = \langle y, (G_N^\epsilon)^\dagger G_N (G_N^\epsilon)^\dagger y \rangle = \langle y, (G_N^\epsilon)^\dagger y \rangle = \sum_{\sigma_n > \epsilon} \frac{|\langle y, v_n \rangle|^2}{\sigma_n}.$$

This gives  $\|\mathcal{T}_N(G_N^\epsilon)^\dagger y\|^2 \leq 1/\min\{\sigma_n : \sigma_n > \epsilon\} \|y\|^2$ , and the result follows.  $\square$

**Remark 5.6** Theorems 5.3 and 5.4 assert convergence to within  $\sqrt{\epsilon}$  only. Using spectral theory techniques it can be shown that  $a^{N,\epsilon} \rightarrow a$  as  $N \rightarrow \infty$  [46, Thm. 5.17], i.e. the regularized coefficients converge to the frame coefficients in the canonical dual frame. Since  $\|f - \mathcal{P}_N^\epsilon f\| = \|\mathcal{T}(a - a^{N,\epsilon})\| \leq$

$\sqrt{B}\|a - a^{N,\epsilon}\|$  this also gives  $\mathcal{P}_N^\epsilon f \rightarrow f$ , albeit at a rate that can be arbitrarily slow. Of course, in finite precision calculations convergence beyond  $\mathcal{O}(\sqrt{\epsilon})$  will not be expected, due to the condition number of the mapping (Theorem 5.5). We note also that convergence down to  $\mathcal{O}(\sqrt{\epsilon})$  has previously been observed empirically in [46, 70]. The main contribution of Theorems 5.3 and 5.4 is that they provide explicit bounds which can be used to estimate the rate of decay of these errors in the regime where they are larger than  $\sqrt{\epsilon}$ .

**Remark 5.7** The approach described in this section has some similarities to standard regularization of ill-posed problems (see, for example, [40, 45, 64]), but also some key differences. First, note that the approximation we seek to compute, i.e.  $\mathcal{P}_N f$ , is a well-conditioned mapping, since  $\mathcal{P}_N$  is an orthogonal projection, whereas the mapping  $f \mapsto x = G_N^{-1} \mathcal{T}_N^* f$ , where  $x \in \mathbb{C}^N$  are the coefficients of  $\mathcal{P}_N$  is ill-conditioned. This ill-conditioned linear system is regularized via truncated SVD, which is a standard approach in ill-posed problems (we note in passing that Tikhonov regularization or its various generalizations could also be used instead, with only minor changes in Theorems 5.3–5.5). The resulting regularized projection  $\mathcal{P}_N^\epsilon$  satisfies the following bounds

$$\|\mathcal{P}_N^\epsilon\| \leq 1, \quad \|(\mathcal{P}_N - \mathcal{P}_N^\epsilon)\mathcal{T}_N\| \leq \sqrt{\epsilon},$$

(these bounds were used implicitly in the proof of Theorem 5.3), which are analogous to standard estimates in the theory of regularization of ill-posed problems (see, for example, [64, eqn. (3)]). Note that the operator  $\mathcal{T}_N$  acts like a ‘smoothing’ operator, in the sense that when  $g = \mathcal{T}_N z$  arises from coefficients  $z \in \mathbb{C}^N$  with  $\|z\| \ll \infty$ , then the projections  $\mathcal{P}_N g$  and  $\mathcal{P}_N^\epsilon g$  are guaranteed to be within  $\mathcal{O}(\sqrt{\epsilon})$  of each other. A key difference between this setting and standard regularization theory is that we are not overly concerned with how well  $x^\epsilon$  solves the linear system  $G_N x = y$ : our interest lies with how well the regularized projection  $\mathcal{P}_N^\epsilon$  approximates the true projection  $\mathcal{P}_N$ . Indeed, since  $\|x\|$  tends to diverge with  $N$ , whereas  $x^\epsilon$  converges to the frame coefficients (Theorem 5.4), we do not expect  $x^\epsilon$  to approximate  $x$  in any sense as  $N \rightarrow \infty$ .

As discussed, ill-conditioning of the discrete problem arises from the ill-posedness of the infinite problem  $\mathcal{G}x = y$ . A standard regularization of this problem involves formulating the least-norm solution, i.e.  $x = \mathcal{G}^\dagger y$ , which is precisely the frame coefficients, i.e.  $x = a = \mathcal{T}^* \mathcal{S}^{-1} f$ . The operator  $\mathcal{S}$  is positive and invertible (and therefore inverting  $\mathcal{S}$  is a well-posed problem), and so it comes as little surprise that stable approximations of the first  $N$  frame coefficients can be computed. However, as discussed, these coefficients generally give poor approximations to  $f$  (see §2.3).

## 5.4 Discussion and Examples

We now consider Theorems 5.3–5.5 in relation to the examples of §3. We focus on two issues: (i) the convergence of the projection  $\mathcal{P}_N^\epsilon f$ , and (ii) the behaviour of the coefficients  $x^\epsilon$ .

First, notice that for small  $N$  – specifically, for  $N$  such that  $\sigma_{\min}(G_N) \geq \epsilon$  – we have  $\mathcal{P}_N^\epsilon = \mathcal{P}_N$ . Hence, the truncated SVD projection initially behaves like the exact projection  $\mathcal{P}_N$ . However, beyond this point the convergence begins to differ. If  $x \in \mathbb{C}^N$  are the coefficients of  $\mathcal{P}_N f$  then setting  $z = x$  in (5.6) gives

$$\|f - \mathcal{P}_N^\epsilon f\| \leq \|f - \mathcal{P}_N f\| + \sqrt{\epsilon}\|x\|.$$

As discussed in §5.1, the term  $\|x\|$  often grows rapidly in  $N$ . Hence as  $N$  increases, the right-hand side of the above inequality may begin to diverge.

However, since  $\Phi$  is a frame there are infinitely-many sequences of coefficients  $c \in \ell^2(I)$  such that  $f = \mathcal{T}c$ . Suppose that  $z = \{c_n\}_{n \in I_N}$ . Then Theorem 5.3 gives

$$\|f - \mathcal{P}_N^\epsilon f\| \leq \|f - \mathcal{T}_N c\| + \sqrt{\epsilon}\|c\|. \quad (5.14)$$

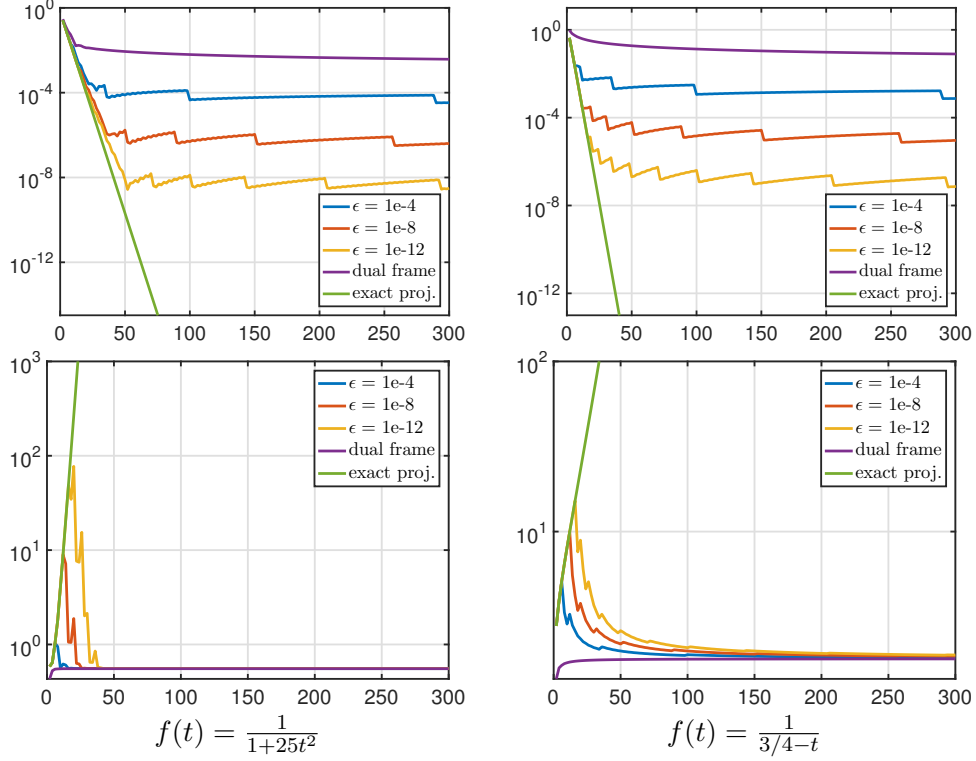


Figure 3: Comparison of the exact projection  $\mathcal{P}_N f$ , the truncated SVD projections  $\mathcal{P}_N^\epsilon f$  and the canonical dual frame expansion  $\sum_{n \in I_N} a_n \phi_n$  for Example 1 with  $T = 2$ . Top row: the errors versus  $N$ . Bottom row: the norms of the coefficient vectors versus  $N$ .

Now the term  $\sqrt{\epsilon}\|c\|$  is independent of  $N$ , while the other term  $\|f - \mathcal{T}_N c\| = \|(\mathcal{T} - \mathcal{T}_N)c\|$  tends to zero as  $N \rightarrow \infty$ . Hence, the rate of decay of the error down to  $\sqrt{\epsilon}$  depends on how well  $f$  can be represented in the frame  $\Phi$  by expansions having small-norm coefficients  $c$ . In the examples below, we show that there always exist coefficient sequences  $c$  that achieve favourable rates of decay of the term  $\|f - \mathcal{T}_N c\|$ . Hence, although the truncated SVD projection  $\mathcal{P}_N^\epsilon f$  may not achieve the same error decay as the exact projection  $\mathcal{P}_N f$ , we can often expect good accuracy.

**Example 1.** A detailed analysis of this example in the one-dimensional case was presented in [9]. Therein it was shown that the projection  $\mathcal{P}_N f$  converges geometrically fast to  $f$  when  $f$  is analytic on  $[-1, 1]$ . However, this is often accompanied by geometric growth of the coefficients  $x$ . Thus, after a certain point, the error of the truncated SVD projection  $\mathcal{P}_N^\epsilon f$  will begin to decay like that of  $\mathcal{P}_N f$ . However, we have the following (see the supplementary material for a proof):

**Proposition 5.8.** *Let  $\Omega \subseteq (-1, 1)^d$  be Lipschitz and consider the frame (3.1). If  $f \in \mathcal{H}^{kd}(\Omega)$  then there exists a set of coefficients  $c \in \ell^2(I)$  such that*

$$\|f - \mathcal{T}_N c\| \leq C_{k,d} N^{-k} \|f\|_{\mathcal{H}^{kd}(-1,1)^d}, \quad \|c\| \leq C_{k,d} \|f\|_{\mathcal{H}^{kd}(-1,1)^d},$$

where  $C_{k,d} > 0$  is independent of  $f$  and  $N$ . In particular, for the exact projection

$$\|f - \mathcal{P}_N f\| \leq C_{k,d} N^{-k} \|f\|_{\mathcal{H}^{kd}(-1,1)^d},$$

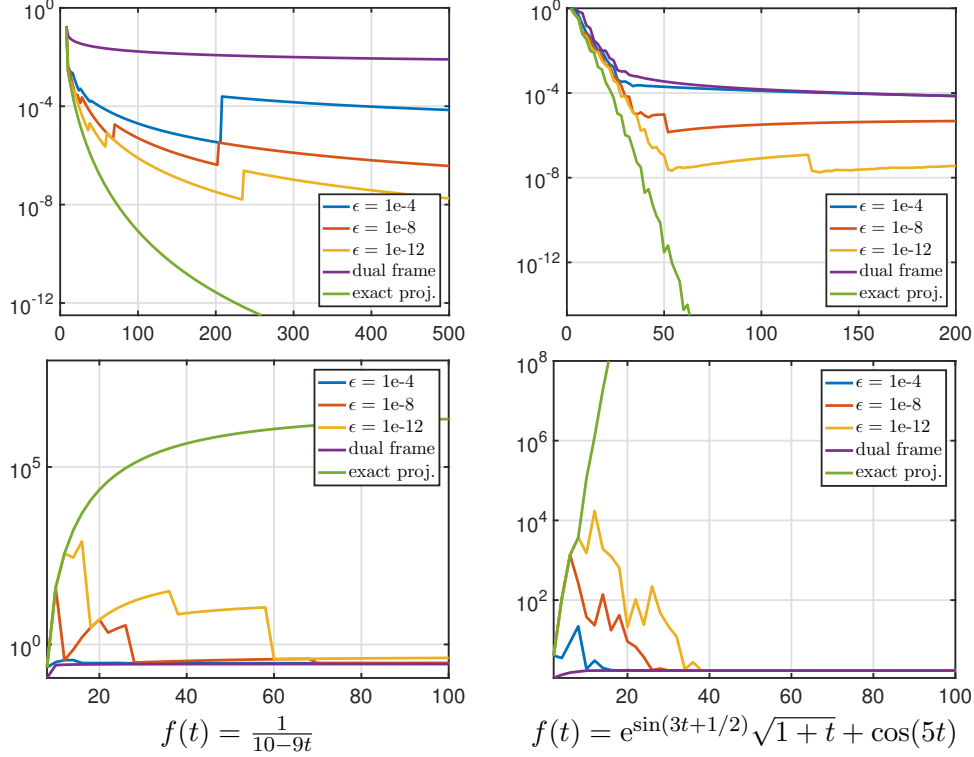


Figure 4: Comparison of the exact projection  $\mathcal{P}_N f$ , the truncated SVD projections  $\mathcal{P}_N^\epsilon f$  and the canonical dual frame expansion  $\sum_{n \in I_N} a_n \phi_n$  for Example 2 with  $K = 8$  (left) and Example 3 with  $\alpha = 1/2$  (right). Top row: the errors versus  $N$ . Bottom row: the norms of the coefficient vectors versus  $N$ .

whereas for the regularized projection,

$$\|f - \mathcal{P}_N^\epsilon f\| \leq C_{k,d} \left( N^{-k} + \sqrt{\epsilon} \right) \|f\|_{H^k(-1,1)^d}. \quad (5.15)$$

This proposition, which holds for arbitrary dimension  $d$ , asserts that there are bounded coefficients vectors for which the error  $\|f - \mathcal{T}_N c\|$  decays at an arbitrarily-fast algebraic rate. In particular, the error of the exact projection decays spectrally fast in  $N$ . On the other hand, for the regularized projection we conclude the following. First, for smooth functions  $f$ , the error decays rapidly in  $N$  when  $\|f - \mathcal{P}_N^\epsilon f\| \gg \sqrt{\epsilon}$ . Second, if the  $k^{\text{th}}$  derivatives of  $f$  grow rapidly in  $k$  then the rate of error decay may lessen as  $\|f - \mathcal{P}_N^\epsilon f\|$  approaches  $\sqrt{\epsilon}$ . This is shown in Figure 3. Note that the derivatives of the function  $f(t) = \frac{1}{1+25t^2}$  grow slowly with  $k$ , whereas the derivatives of  $f(t) = \frac{1}{3/4-t}$  grow much more rapidly. As predicted by (5.15), there is substantially less effect from replacing  $\mathcal{P}_N$  with the regularized projection  $\mathcal{P}_N^\epsilon$  in the case of the former than in the latter. This aside, Fig. 3 also shows the slow convergence of the canonical dual frame expansion (2.9), thus confirming the discussion in §3, and the norms of the various coefficient vectors. These results are in good agreement with Theorem 5.4: the coefficients of  $\mathcal{P}_N^\epsilon f$  initially grow large, but after a certain point they begin to decay to the limiting value  $\|a\|$ .

**Example 2.** In this case we observe algebraic convergence at a rate determined by the number of polynomials added to the Fourier basis:



**Proposition 5.9.** *Let  $K \in \mathbb{N}$  be fixed and consider the frame (3.3). If  $f \in H^k(-1, 1)$  for  $0 \leq k \leq K$  then there exists a set of coefficients  $c \in \ell^2(I)$  such that*

$$\|f - \mathcal{T}_N w\| \leq C_k N^{-k} \|f\|_{H^k(-1,1)}, \quad \|c\| \leq C_k \|f\|_{H^k(-1,1)},$$

where  $C_k > 0$  is independent of  $f$  and  $N$ . In particular,

$$\|f - \mathcal{P}_N^\epsilon f\| \leq C_k N^{-k} \|f\|_{H^k(-1,1)^d},$$

and

$$\|f - \mathcal{P}_N^\epsilon f\| \leq C_k \left( N^{-k} + \sqrt{\epsilon} \right) \|f\|_{H^k(-1,1)^d}.$$

This result establishes algebraic convergence in this frame. See Fig. 4.

**Example 3.** We have:

**Proposition 5.10.** *Let  $\mathcal{Q}_N : L^2(-1, 1) \rightarrow L^2(-1, 1)$  be the orthogonal projection onto  $\mathbb{P}_{N/2-1}$ . If  $f(t) = w(t)g(t) + h(t)$ , then there exists a vector  $c \in \ell^2(I)$  such that*

$$\|f - \mathcal{T}_N w\| \leq w_{\max} \|g - \mathcal{Q}_N g\| + \|h - \mathcal{Q}_N h\|, \quad \|c\| \leq \|g\| + \|h\|,$$

where  $w_{\max} = \text{ess sup}_{t \in (-1,1)} |w(t)|$ . In particular,

$$\|f - \mathcal{P}_N f\| \leq w_{\max} \|g - \mathcal{Q}_N g\| + \|h - \mathcal{Q}_N h\|,$$

and

$$\|f - \mathcal{P}_N^\epsilon f\| \leq w_{\max} \|g - \mathcal{Q}_N g\| + \|h - \mathcal{Q}_N h\| + \sqrt{\epsilon} (\|g\| + \|h\|).$$

This result implies the convergence of the regularized projection  $\mathcal{P}_N^\epsilon f$  is spectral in the factors  $g$  and  $h$ . In particular, if  $g$  and  $h$  are smooth then one sees superalgebraic convergence down to  $\sqrt{\epsilon}$ , and if  $g$  and  $h$  are analytic, then one has geometric convergence down to  $\sqrt{\epsilon}$ . Unlike the previous two examples, there is no lessening of the error decay when  $g$  or  $h$  have large derivatives. This is shown in Fig. 4. This figure suggests geometric convergence, in agreement with Proposition 5.10, but with a somewhat reduced exponent over that of the exact projection  $\mathcal{P}_N f$ . In other words, there exist coefficient vectors in the frame which yield faster geometric convergence, but are too large in norm to be obtained as solutions of the regularized system.

## 6 Numerically stable frame approximations

Up to this point, the accuracy and stability of the numerical frame projection  $\mathcal{P}_N^\epsilon$  are limited to  $\mathcal{O}(\sqrt{\epsilon})$  and  $\mathcal{O}(1/\sqrt{\epsilon})$  respectively. We close this paper with a brief description of an approximation that is stable and achieves  $\mathcal{O}(\epsilon)$  accuracy. This is based on oversampling. Specifically, rather than solving the square system (2.16), we consider the  $M \times N$  system

$$G_{M,N} x \approx y, \quad y = \{\langle f, \phi_n \rangle\}_{n \in I_M}, \quad (6.1)$$

where  $G_{M,N} = \{\langle \phi_n, \phi_m \rangle\}_{m \in I_M, n \in I_N} \in \mathbb{C}^{M \times N}$ . Note that where  $G_N$  corresponds to the finite section of the Gram operator  $\mathcal{G}$ ,  $G_{M,N}$  corresponds to a so-called *uneven section*. Uneven sections are known to be useful alternatives to finite sections in computational spectral theory [14, 41, 43, 44, 57] and, more recently, sampling theory [3, 4, 5, 13]. Much the same is true in this instance.

Since  $\mathcal{G}$  is singular, then matrix  $G_{M,N}$  remains ill-conditioned even when  $M \geq N$ . Hence we consider the regularized solution and corresponding projection

$$x^\epsilon = (G_{M,N}^\epsilon)^\dagger y, \quad \mathcal{P}_{M,N}^\epsilon f = \sum_{n \in I_N} (x^\epsilon)_n \phi_n.$$

where  $G_{M,N}^\epsilon$  is obtained by discarding all its singular values of  $G_{M,N}$  below  $\epsilon$ . We now claim that, given sufficient oversampling, the projection  $\mathcal{P}_{M,N}^\epsilon$  is stable and achieves  $\mathcal{O}(\epsilon)$  accuracy. To this end, we define the following two constants:

$$\kappa_{M,N}^\epsilon = \max_{\substack{y \in \mathbb{C}^M \\ \|y\|=1}} \left\| \mathcal{T}_N (G_{M,N}^\epsilon)^\dagger \right\|, \quad \lambda_{M,N}^\epsilon = \epsilon^{-1} \max_{\substack{z \in \mathbb{C}^N \\ \|z\|=1}} \left\| \mathcal{T}_N z - \mathcal{P}_{M,N}^\epsilon \mathcal{T}_N z \right\|.$$

We now have the following generalization of Theorems 5.3 and 5.4:

**Theorem 6.1.** *The truncated SVD projection  $\mathcal{P}_{M,N}^\epsilon$  satisfies*

$$\|f - \mathcal{P}_{M,N}^\epsilon f\| \leq \left(1 + \sqrt{B} \kappa_{M,N}^\epsilon\right) \|f - \mathcal{T}_N z\| + \epsilon \lambda_{M,N}^\epsilon \|z\|, \quad \forall z \in \mathbb{C}^N,$$

Moreover, the coefficients satisfy

$$\|x^\epsilon\| \leq \sqrt{B}/\epsilon \|f - \mathcal{T}_N z\| + \|z\|, \quad \forall z \in \mathbb{C}^N,$$

and if  $a^{M,N,\epsilon} \in \ell^2(I)$  is the extension of  $x^\epsilon$  by zero,

$$\|a - a^{M,N,\epsilon}\| \leq (1 + B/\epsilon) \sqrt{\sum_{n \in I \setminus I_N} |a_n|^2} + \epsilon \lambda_{M,N}^\epsilon / \sqrt{A} \|a\|.$$

This result establishes the above claim, provided the constants  $\kappa_{M,N}^\epsilon$  and  $\lambda_{M,N}^\epsilon$  are  $\mathcal{O}(1)$  as  $\epsilon \rightarrow 0$ . For this, we note the following:

**Proposition 6.2.** *For fixed  $\epsilon > 0$  and  $N \in \mathbb{N}$ , the constants  $\kappa_{M,N}^\epsilon$  and  $\lambda_{M,N}^\epsilon$*

$$\limsup_{M \rightarrow \infty} \kappa_{M,N}^\epsilon \leq 1/\sqrt{A}, \quad \limsup_{M \rightarrow \infty} \lambda_{M,N}^\epsilon \leq 1/\sqrt{A}.$$

In particular, as long as  $M$  is chosen sufficiently large, one has the error bound

$$\|f - \mathcal{P}_{M,N}^\epsilon f\| \leq C (\|f - \mathcal{T}_N z\| + \epsilon \|z\|), \quad \forall z \in \mathbb{C}^N,$$

for some  $C > 0$ . This is identical to the error bound (5.6) for the projection  $\mathcal{P}_N^\epsilon f$ , except for appearance of  $\epsilon$  in place of  $\sqrt{\epsilon}$ . Hence the error now decays down to  $\mathcal{O}(\epsilon)$ , with, as in the case of  $\mathcal{P}_N^\epsilon f$ , the rate of decay of the error being dictated by the convergence rate of expansions in the frame with small-norm coefficients.

We defer the proofs of these results for a follow-up work [8], as they are part of a much more general topic on frame approximations from ‘indirect’ data. Note that this work includes discrete function samples, for example, which are typically much more convenient to work with than (6.1) since they do not require evaluations of inner products. Instead, in Fig. 5 we present several numerical results for the Examples 1–3 based on (6.1). These results illustrate that with a reasonably small amount of oversampling, e.g.  $M = 2N$ , one can obtain a much more accurate numerical frame approximation than when  $M = N$ , which is the case considered previously (see [10] for further analysis in the case of Example 1). For these cases, similar results can also be obtained via least-squares fitting with discrete function samples taken, for example, on an equally-spaced grid of  $M$  points [8].

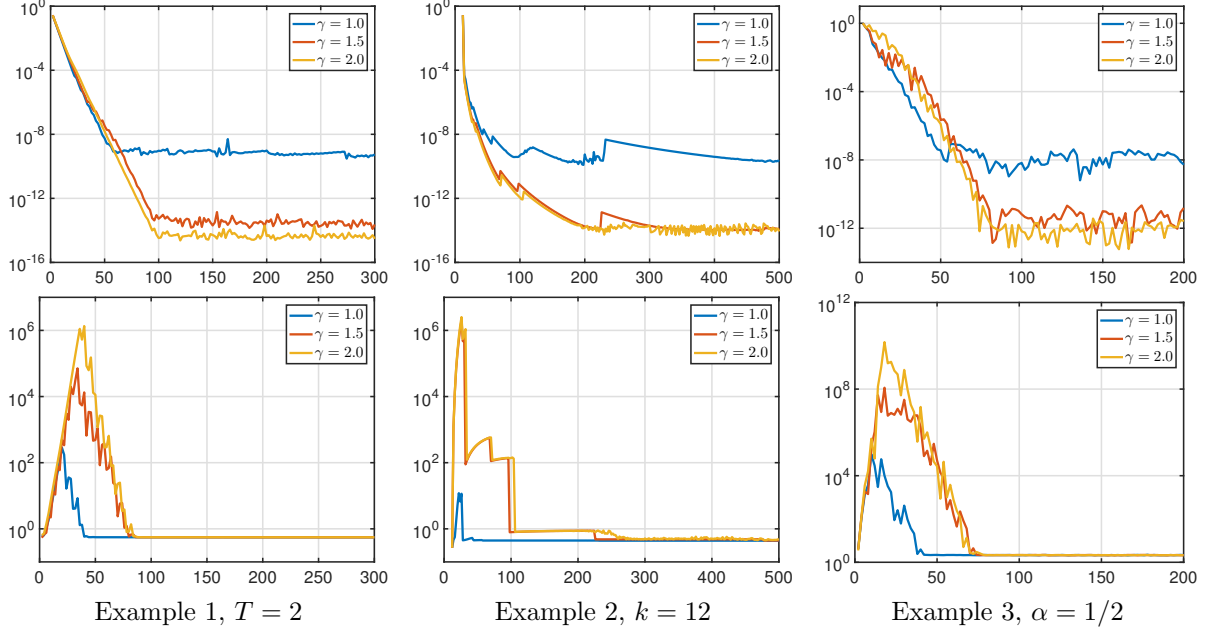


Figure 5: Comparison of the projection  $\mathcal{P}_{\gamma N, N}^\epsilon f$  for different values of the parameter  $\gamma$ . Top row: the errors versus  $N$ . Bottom row: the norms of the coefficient vectors versus  $N$ . The functions used were  $f(t) = \frac{1}{1+25t^2}$  (left),  $f(t) = \frac{1}{5-4t}$  (middle) and  $f(t) = e^{\sin(3t+1/2)}\sqrt{1+t} + \cos(5t)$  (right).

## 7 Conclusions and further research

The concern of this paper has been computing numerical approximations in frames, in particular, orthogonal projections in the span of the first  $N$  frame elements. There are four main conclusions. First, truncated frames always lead to ill-conditioned linear systems. As shown, such ill-conditioning is due to approximation of the singular Gram operator by its finite section, and can be arbitrarily bad depending on the frame. Second, the orthogonal projection typically cannot be computed in practice, since it generally has coefficients that grow rapidly in  $N$ . However, by regularization it is possible to compute a numerical approximation in a truncated frame which has bounded instability (up to  $\mathcal{O}(1/\sqrt{\epsilon})$ ) and which converges down to an error of  $\mathcal{O}(\sqrt{\epsilon})$ . Moreover, the convergence of this approximation depends on how well  $f$  can be approximated by finite expansions in the frame with small norm coefficients. Fourth, by oversampling, one can compute a truly stable approximation that is convergent down to  $\mathcal{O}(\epsilon)$ .

The overall conclusion of this paper is that satisfactory approximations can be computed in certain finite systems with near-linear dependencies. We stress that the frame condition is crucial in this regard. The monomials  $\Phi_N = \{1, x, x^2, \dots, x^{N-1}\}$  are nearly-linearly dependent for large  $N$ , but do not lead to good numerical approximations since smooth functions do not necessarily have approximations in this system with small-norm coefficients. On the other hand, a frame guarantees at least one approximation with small-norm coefficients, namely, the truncated canonical dual frame expansion, although, as seen, better approximations often exist.

There a number of topics not considered in this paper. First is a more detailed analysis of the least-squares frame approximations, considered briefly in §6. This also includes the case of approximations from ‘indirect’ data, e.g. pointwise samples. This will be addressed in detail in the follow-up work [8]. Second, since the focus of this paper has been on general frames, we have not

considered fast computations (which is more specific to the frame employed). A fast algorithm for computing Fourier extensions in the one-dimensional setting was introduced in [60], for the special case where the extension interval has exactly twice the length of the original interval. A more recent alternative that also generalizes to higher dimensions is described in [63]. Crucial elements in the latter approach include the link to the theory of bandlimited functions, the special prolate spheroidal wave functions and a phenomenon called the *plunge region* in sampling theory. It seems these elements may generalize to other types of frames beyond Fourier extensions, which is a topic that will be considered in future work. Third and finally, we have not discussed the accuracy of computing the SVD of the Gram matrix, and its effect on the numerical projection. Numerical experiments suggests this does not have a substantial effect on the approximation error, but since the Gram matrix is severely ill-conditioned a careful analysis should be carried out. We expect the structure of the singular values (in particular, their tendency to divide into ‘good’ singular value away from zero and ‘bad’ singular values near zero) is important in this regard.

## Acknowledgements

The initial ideas for this paper were first discussed during the Research Cluster on “Computational Challenges in Sparse and Redundant Representations” at ICERM in November 2014. The authors would like to thank all the participants for the useful discussions and feedback received during the program. They would also like to thank John Benedetto, Pete Casazza, Vincent Coppé, Roel Matthysen, Nick Trefethen, Mikael Slevinsky, Thomas Strohmer, Andy Wathen and Marcus Webb.

The first author is supported by NSERC grant 611675, as well as an Alfred P. Sloan Research Fellowship. The second author is supported by FWO-Flanders projects G.0641.11 and G.A004.14, as well as by KU Leuven project C14/15/055.

## References

- [1] B. Adcock. *Modified Fourier expansions: theory, construction and applications*. PhD thesis, University of Cambridge, 2010.
- [2] B. Adcock. Convergence acceleration of modified Fourier series in one or more dimensions. *Math. Comp.*, 80(273):225–261, 2011.
- [3] B. Adcock and A. C. Hansen. A generalized sampling theorem for stable reconstructions in arbitrary bases. *J. Fourier Anal. Appl.*, 18(4):685–716, 2012.
- [4] B. Adcock and A. C. Hansen. Generalized sampling and infinite-dimensional compressed sensing. *Found. Comput. Math.*, 16(5):1263–1323, 2016.
- [5] B. Adcock, A. C. Hansen, and C. Poon. Beyond consistent reconstructions: optimality and sharp bounds for generalized sampling, and application to the uniform resampling problem. *SIAM J. Math. Anal.*, 45(5):3114–3131, 2013.
- [6] B. Adcock and D. Huybrechs. On the resolution power of Fourier extensions for oscillatory functions. *J. Comput. Appl. Math.*, 260:312–336, 2014.
- [7] B. Adcock and D. Huybrechs. Frames and numerical approximation – supplementary material. *Preprint*, 2016.
- [8] B. Adcock and D. Huybrechs. Frames and stable numerical approximation from indirect data. *In preparation*, 2016.
- [9] B. Adcock, D. Huybrechs, and J. Martín-Vaquero. On the numerical stability of Fourier extensions. *Found. Comput. Math.*, 14(4):635–687, 2014.

- [10] B. Adcock and J. Ruan. Parameter selection and numerical approximation properties of Fourier extensions from fixed data. *J. Comput. Phys.*, 273:453–471, 2014.
- [11] N. Albin and O. P. Bruno. A spectral FC solver for the compressible Navier–Stokes equations in general domains I: Explicit time-stepping. *J. Comput. Phys.*, 230(16):6248–6270, 2011.
- [12] J. J. Benedetto. Irregular sampling and frames. In C. K. Chui, editor, *Wavelets: A Tutorial in Theory and Applications*. Boca Raton, FL: CRC, 1994.
- [13] P. Berger and K. Gröchenig. Sampling and reconstruction in different subspaces by using oblique projections. *arXiv:1312.1717*, 2013.
- [14] A. Böttcher. Infinite matrices and projection methods. In *Lectures on operator theory and its applications (Waterloo, ON, 1994)*, volume 3 of *Fields Inst. Monogr.*, pages 1–72. Amer. Math. Soc., Providence, RI, 1996.
- [15] J. Boyd. Fourier embedded domain methods: extending a function defined on an irregular region to a rectangle so that the extension is spatially periodic and  $C^\infty$ . *Appl. Math. Comput.*, 161(2):591–597, 2005.
- [16] J. P. Boyd. A comparison of numerical algorithms for Fourier Extension of the first, second, and third kinds. *J. Comput. Phys.*, 178:118–160, 2002.
- [17] O. Bruno and M. Lyon. High-order unconditionally stable FC-AD solvers for general smooth domains I. Basic elements. *J. Comput. Phys.*, 229(6):2009–2033, 2010.
- [18] O. P. Bruno, Y. Han, and M. M. Pohlman. Accurate, high-order representation of complex three-dimensional surfaces via Fourier continuation analysis. *J. Comput. Phys.*, 227(2):1094–1125, 2007.
- [19] E. J. Candès and D. Donoho. Ridgelets: a key to higher-dimensional intermittency? *Phil. Trans. R. Soc. Lond. A*, 357(10):2495–2509, 1999.
- [20] E. J. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise  $C^2$  singularities. *Comm. Pure Appl. Math.*, 57(2):219–266, 2004.
- [21] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Appl. Comput. Harmon. Anal.*, 31(1):59–73, 2010.
- [22] P. G. Casazza. The art of frame theory. *Taiwanese J. Math.*, 4(2):129–202, 2000.
- [23] P. G. Casazza and O. Christensen. Approximation of the frame coefficients using finite-dimensional methods. *J. Electronic Imaging*, 06(04):479–483, 1997.
- [24] P. G. Casazza and O. Christensen. Riesz frames and approximation of the frame coefficients. *Approx. Theory Appl.*, 14(2):1–11, 1998.
- [25] P. G. Casazza and O. Christensen. Approximation of the inverse frame operator and applications to gabor frames. *J. Approx. Theory*, 103:338–356, 2000.
- [26] P. G. Casazza and G. Kutyniok, editors. *Finite Frames: Theory and Applications*. Birkhauser, 2013.
- [27] Z. Chen and C.-W. Shu. Recovering exponential accuracy from collocation point values of smooth functions with end-point singularities. *J. Comput. Appl. Math.*, 265:83–95, 2014.
- [28] Z. Chen and C.-W. Shu. Recovering exponential accuracy in fourier spectral methods involving piecewise smooth functions with unbounded derivative singularities. *Preprint*, 2014.
- [29] O. Christensen. Frames and the projection method. *Appl. Comput. Harmon. Anal.*, 1:50–53, 1993.
- [30] O. Christensen. Frames containing a Riesz basis and approximation of the frame coefficients using finite dimensional methods. *J. Math. Anal. Appl.*, 199:256–270, 1996.
- [31] O. Christensen. Finite-dimensional approximation of the inverse frame operator and applications to Weyl–Heisenberg frames and wavelet frames. *J. Fourier Anal. Appl.*, 6:79–91, 200.

- [32] O. Christensen. *An Introduction to Frames and Riesz Bases*. Birkhauser, 2003.
- [33] O. Christensen and A. Lindner. Frames containing a Riesz basis and approximation of the inverse frame operator. *Internat. Ser. Numer. Math.*, 137:89–100, 2001.
- [34] O. Christensen and T. Strohmer. The finite section method and problems in frame theory. *J. Approx. Theory*, 133:221–237, 2005.
- [35] I. Daubechies. *Ten lectures on wavelets*. SIAM, Philadelphia, 1992.
- [36] I. Daubechies, A. Grossmann, and Y. Meyer. Painless nonorthogonal expansions. *J. Math. Phys.*, pages 1271–1283, 1986.
- [37] E. B. Davies and M. Plum. Spectral pollution. *IMA J. Num. Anal.*, 23(3):417–438, 2004.
- [38] R. J. Duffin and A. C. Schaeffer. A class of nonharmonic Fourier series. *Trans. Amer. Math. Soc.*, 72(2):341–366, 1952.
- [39] K. S. Eckhoff. Accurate and efficient reconstruction of discontinuous functions from truncated series expansions. *Math. Comp.*, 61(204):745–763, 1993.
- [40] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic, Dordrecht, The Netherlands, 1996.
- [41] K. Gröchenig, Z. Rzesotnik, and T. Strohmer. Quantitative estimates for the finite section method and Banach algebras of matrices. *Integral Equations and Operator Theory*, 67(2):183–202, 2011.
- [42] C. W. Groetsch. *Generalized Inverses of Linear Operators: Representation and Approximation*. Marcel Dekker Inc., 1977.
- [43] A. C. Hansen. On the approximation of spectra of linear operators on Hilbert spaces. *J. Funct. Anal.*, 254(8):2092–2126, 2008.
- [44] A. C. Hansen. On the solvability complexity index, the n-pseudospectrum and approximations of spectra of operators. *J. Amer. Math. Soc.*, 24(1):81–124, 2011.
- [45] P. C. Hansen, V. Pereyra, and G. Scherer. *Least Squares Data Fitting with Applications*. John Hopkins University Press, Baltimore, 2012.
- [46] M. L. Harrison. *Frames and irregular sampling from a computational perspective*. PhD thesis, University of Maryland – College Park, 1998.
- [47] D. P. Hewett, S. N. Chandler-Wilde, S. Langdon, and A. Twigger. A high frequency boundary element method for scattering by a class of nonconvex obstacles. *Numer. Math.*, 129:647–689, 2015.
- [48] J. A. Hogan and J. D. Lakey. *Duration and Bandwidth Limiting*. Birkhäuser, 2012.
- [49] D. Huybrechs. On the Fourier extension of non-periodic functions. *SIAM J. Numer. Anal.*, 47(6):4326–4355, 2010.
- [50] M. Javed and L. N. Trefethen. Euler-Maclaurin and Gregory interpolants. *Numer Math*, 132:201–216, 2016.
- [51] J. Kovacevic and A. Chebira. Life beyond bases: The advent of frames (part 2). *IEEE Signal Process. Mag.*, 24(5):115–125, 2007.
- [52] J. Kovacevic and A. Chebira. Life beyond bases: The advent of frames (part i). *IEEE Signal Process. Mag.*, 24(4):86–104, 2007.
- [53] A. Krylov. On approximate calculations. *Lectures delivered in 1906 (in Russian)*. St Petersburg, 1907.
- [54] G. Kutyniok and D. Labate, editors. *Shearlets: Multiscale Analysis for Multivariate Data*. Springer, 2012.
- [55] H. J. Landau and H. O. Pollak. Prolate spheroidal wave functions, Fourier analysis and uncertainty—III: The dimension of the space of essentially timeand bandlimited signals. *Bell System Tech J.*, 41(4):1295–1336, 1962.

- [56] M. Lewin and É. Séré. Spectral pollution and how to avoid it. *Proc. London Math. Soc.*, 100(3):864–900, 2009.
- [57] M. Lindner. *Infinite Matrices and their Finite Sections*. Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2006.
- [58] S. H. Lui. Spectral domain embedding for elliptic PDEs in complex domains. *J. Comput. Appl. Math.*, 225(2):541–557, 2009.
- [59] M. Lyon. Approximation error in regularized SVD-based Fourier continuations. *Appl. Numer. Math.*, 62:1790–1803, 2012.
- [60] M. Lyon. A fast algorithm for Fourier continuation. *SIAM J. Sci. Comput.*, 33(6):3241–3260, 2012.
- [61] M. Lyon and O. Bruno. High-order unconditionally stable FC-AD solvers for general smooth domains II. Elliptic, parabolic and hyperbolic PDEs; theoretical considerations. *J. Comput. Phys.*, 229(9):3358–3381, 2010.
- [62] S. G. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 3 edition, 2009.
- [63] R. Matthysen and D. Huybrechs. Fast algorithms for the computation of Fourier extensions of arbitrary length. *SIAM J. Sci. Comput.*, 2015. To appear.
- [64] A. Neumaier. Solving ill-conditioned and singular linear systems: a tutorial on regularization. *SIAM Rev.*, 40(3):636–666, 1998.
- [65] R. Pasquetti and M. Elghaoui. A spectral embedding method applied to the advection–diffusion equation. *J. Comput. Phys.*, 125:464–476, 1996.
- [66] R. Platte, A. J. Gutierrez, and A. Gelb. Edge informed Fourier reconstruction from non-uniform spectral data with exponential convergence rates. *Preprint*, 2012.
- [67] D. Slepian. Prolate spheroidal wave functions. Fourier analysis, and uncertainty V: The discrete case. *Bell System Tech J.*, 57:1371–1430, 1978.
- [68] G. Song and A. Gelb. Approximating the inverse frame operator from localized frames. *Appl. Comput. Harmon. Anal.*, 35:94–110, 2013.
- [69] D. B. Stein, R. D. Guy, and B. Thomases. Immersed boundary smooth extension: a high-order method for solving PDE on arbitrary smooth domains using Fourier spectral methods. *J. Comput. Phys.*, 304:252–274, 2016.
- [70] T. Strohmer. Numerical analysis of the nonuniform sampling problem. *J. Comput. Appl. Math.*, 122:297–316, 2000.
- [71] A. Teolis and J. J. Benedetto. Local frames and noise reduction. *Signal Process.*, 45:369–387, 1995.
- [72] L. N. Trefethen. *Approximation theory and approximation practice*. SIAM, Philadelphia, 2013.